FACTORS UNDERLYING RAPID REPRODUCTIVE PROTEIN EVOLUTION IN DROSOPHILA

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

In Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Alex Wong

May 2008

FACTORS UNDERLYING RAPID REPRODUCTIVE PROTEIN EVOLUTION IN DROSOPHILA

Alex Wong, Ph. D.

Cornell University 2008

Biologists have long noted the tremendous diversity of behaviors, morphological traits and molecules involved in mating and reproduction. In this thesis, I investigate the molecular evolution of reproductive proteins in the vinegar fly *Drosophila melanogaster*, focusing on a class of ejaculate proteins known as accessory gland proteins ("Acps"). Previous work has documented extensive evidence for rapid, adaptive evolution of some Acps. It is generally thought that male-female interactions, e.g., sexual conflict and cryptic female choice, drive rapid Acp evolution, although evidence specifically favoring this hypothesis in *D. melanogaster* is limited. Here, I describe biochemical and structural studies on a particularly rapidly evolving Acp, ovulin. I argue that structural features of ovulin may contribute to its ability to tolerate high sequence diversity. I also investigate the molecular evolution of a class of Acps and female reproductive tract proteins that (I argue) are particularly likely to undergo co-evolution between males and females, namely proteolysis regulators and targets of proteolysis. I show that a number of proteolysis regulators and targets are subject to positive selection, and find evidence of male-female co-evolution. Finally, I critically examine an underlying assumption of many divergence based methods for inferring positive selection – the assumption of phylogenetic congruence between loci. I find that, within the genus Drosophila, at least two nodes show evidence for phylogenetic incongruence, possibly due to incomplete ancestral lineage sorting.

BIOGRAPHICAL SKETCH

Alex Wong received a B.A. in Biology and Philosophy from Carleton University in 2000, and a M.A. in Philosophy in 2002 from the same institution. He will go on to pursue his interests in evolutionary genetics in a post-doctoral fellowship at the University of Ottawa. Alex thinks that truth, beauty, and the good are still worth pursuing.

For Amanda and Myles.

ACKNOWLEDGMENTS

The work described in this thesis would not have been possible without the help and input of many friends and colleagues. Mariana Wolfner and Chip Aquadro have been excellent mentors, and I thank them for their intellectual rigor, commitment to science, and personal warmth. My committee members Rick Harrison and Andrew Clark have provided keen insight, strong support, and a critical eye. Jeff Jensen, Laura Sirot, Ravi Ram Kristipati, Lisa McGraw, John Pool, Kristi Montooth and José Andrés (to name a few) have been wonderful friends and collaborators. Funding for the work described in this thesis comes from NIH grants to Charles Aquadro and Mariana Wolfner, a NSF doctoral dissertation improvement grant, and fellowship support from HHMI.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

INTRODUCTION – CAUSES OF RAPID REPRODUCTIVE PROTEIN EVOLUTION: THE VIEW FROM DROSOPHILA

"What is the use and what is the power of semen? Is it to be reckoned as two principles, the material and the active, as Hippocrates supposed, or only one of them, the efficient, as in the opinion of Aristotle, who…does not grant that any part of the animal is formed from it? The matter deserves to be investigated and the disagreement of such great men adjudicated, not by recourse to plausible arguments, but by a demonstration that begins from and proceeds through what is clearly evident."

- Galen, "On Semen", ~160 AD

Introduction

Since the mid-1800s, biologists have noted the tremendous diversity of behaviors, morphological traits and, more recently, molecules involved in mating and reproduction. Charles Darwin developed the theory of sexual selection to explain the seemingly maladaptive nature of many characters involved in the acquisition of mates, such as the peacock's tail (Darwin 1871). He argued that male-male competition and female preferences could account for the evolution of flashy male display traits, since a male's success in procuring mates would substantially increase his contributions to future generations.

Since Darwin's time, sexual selection has become an important component of evolutionary thought. Theoretical and empirical studies have established both the possibility and the reality of sexual selection as an important evolutionary force (e.g., (Andersson 1994; Eberhard 1996; Birkhead and Møller 1998). While such studies have their origins in morphology and behavioral ecology, an increasing body of recent

work has focused on the evolution of reproductive molecules, particularly of proteins involved in gamete recognition and fusion, sperm storage, and the control of mating/post-mating behavior (reviewed in Clark, Aagaard, and Swanson 2006; Panhuis, Clark, and Swanson 2006). Studies in a wide variety of organisms have noted that genes expressed in reproductive tissues (primarily male) tend to evolve rapidly between species, often under the influence of positive selection.

Researchers have generally invoked post-copulatory sexual selection as the most likely explanation for the rapid evolution of genes expressed in reproductive tissues (Swanson and Vacquier 2002; Clark, Aagaard, and Swanson 2006; Panhuis, Clark, and Swanson 2006). By analogy to pre-copulatory mate choice, it is thought that post-copulatory gametic choice can, and does, exert strong selective pressures on reproductive tract proteins. Here, I review the evidence for post-copulatory sexual selection on reproductive tract proteins, with particular reference to *Drosophila*. The available data, I will argue, while consistent to varying degrees with post-copulatory sexual selection, are largely not uniquely predicted by this family of hypotheses. Other possible causes of rapid evolution, particularly host-pathogen interactions, need to be taken into account when considering the molecular evolution of reproductive tract proteins. New data and methods should open up fruitful avenues for narrowing down the range of feasible explanations.

The accessory gland proteins of Drosophila melanogaster

The seminal fluid proteins of *D. melanogaster* have been studied extensively with respect to both function and molecular evolution, and are often cited as likely targets of post-copulatory sexual selection. As such, they provide a good case study in dissecting the roles of natural and sexual selection in the evolution of reproductive tract genes. In Drosophila and many other insects, the male accessory glands are

major contributors to the seminal fluid (for reviews, see Gillott 2003; Chapman and Davies 2004; Ravi Ram and Wolfner 2007). Accessory gland products are best understood in *D. melanogaster*, although recent studies have begun to characterize the seminal fluid of a number of other insects (Andres et al. 2006; Braswell et al. 2006; Collins et al. 2006; Davies and Chapman 2006; Dottorini et al. 2007; Sirot et al. 2008).

Accessory gland proteins (Acps) are required for most or all of the post-mating changes undergone by *D. melanogaster* females following mating (reviewed in Ravi Ram and Wolfner 2007). Mated females actively reject copulation attempts by subsequent males, increase their rates of egg production, ovulation, egg-laying, and feeding, and store sperm for up to two weeks. In the context of multiple mating (a common event in many species of Drosophila, including *D. melanogaster*), sperm storage in turn sets the stage for sperm competition, whereby sperm from different males compete for fertilization opportunities. Moreover, mated *D. melanogaster* females suffer a 'cost of mating', in that their lifespan is reduced in comparison to virgins. Products of the accessory gland are necessary for all of these post-mating responses: Males lacking accessory glands, or whose accessory gland secretory cells have been largely ablated, fail to elicit any of these responses in their mates (Xue and Noll 2000; Kalb, DiBenedetto, and Wolfner 1993; Chapman et al. 1995; Tram and Wolfner 1998).

In *D. melanogaster*, genetic, transgenic, biochemical, and association studies have identified or suggested specific roles for a number of Acps in the female post-mating responses and sperm competition (Table 1.1). For example, the large glycoprotein Acp36DE and the lectin Acp29AB are necessary for female sperm storage (Neubaum and Wolfner 1999; Bloch Qazi and Wolfner 2003; Wong et al. in preparation), and play roles in sperm competition (Clark et al. 1995; Chapman et al.

Table 1.1 Effects of *D. melanogaster* Acps.

Gene	Type	Assay	Phenotype	Reference
Sex-peptide	Peptide	Injection; knockout/down; ectopic expression; association	Egg-laying; refractoriness; toxicity; feeding	(Chen et al. 1988; Wigby and Chapman 2005; Carvalho et al. 2006; Fiumera, Dumont, and Clark 2007)
Ovulin	Peptide	Knockout; ectopic expression; association	Ovulation; P1, P21	(Clark et al. 1995; Herndon and Wolfner 1995; Fiumera, Dumont, and Clark 2005; Heifetz et al. 2005)
CG10433	Peptide	Ectopic expression	Toxicity	(Mueller, Page, and Wolfner 2007)
Acp33A	Peptide	Association	P1, P2	(Fiumera, Dumont, and Clark 2005)
Acp53E	Peptide	Association	P1	(Clark et al. 1995)
CG14560	Peptide	Association	P1, Fecundity	(Fiumera, Dumont, and Clark 2007)
Mst57	Peptide	Association	Fecundity	(Fiumera, Dumont, and Clark 2007)
Acp36DE	Glycoprotein	Knockout, association	Sperm storage; P1; P22	(Clark et al. 1995; Neubaum and Wolfner 1999; Chapman et al. 2000)
Acp29AB	Lectin	Knockout, association	Sperm storage; P1; P22	(Clark et al. 1995; Fiumera, Dumont, and Clark 2005) Wong et al. in preparation)
Acp62F	Protease inhibitor	Knockout, association, ectopic expression, association	P1, P22; fecundity; toxicity; ovulin processing	(Lung et al. 2002; Fiumera, Dumont, and Clark 2007; Mueller et al. 2008)
CG9334	Protease inhibitor	Ectopic expression	Bacterial clearance	(Mueller, Page, and Wolfner 2007)

Table 1.1 (Continued)

CG8137	Protease inhibitor	Ectopic expression; association	Toxicity; P1	(Fiumera, Dumont, and Clark 2005; Mueller, Page, and Wolfner 2007)
CG9997	Protease	Knockdown	Egg-laying; refractoriness; sperm release	(Ram and Wolfner 2007)
CG10284	Protease	Ectopic expression	Bacterial clearance	(Mueller, Page, and Wolfner 2007)
CG6168	Protease	Ectopic expression, association	Bacterial clearance; refractoriness; P2	(Fiumera, Dumont, and Clark 2005; Mueller, Page, and Wolfner 2007)
CG17331	Protease	Association	P2; refractoriness; mortality	(Fiumera, Dumont, and Clark 2005; Fiumera, Dumont, and Clark 2006)
CG11864	Protease	Knockdown	Ovulin processing	(Ravi Ram, Sirot, and Wolfner 2006)
CG32382	Protease	Knockdown	Systemic immune response	(Kambris et al. 2006)
CG32383	Protease	Knockdown	Systemic immune response	(Kambris et al. 2006)
CG1652	Lectin	Knockdown	Egg-laying; refractoriness; sperm release	(Ram and Wolfner 2007)
CG1656	Lectin	Knockdown	Egg-laying; refractoriness; sperm release	(Ram and Wolfner 2007)
CG17575	CRISP	Knockdown	Egg-laying; refractoriness; sperm release	(Ram and Wolfner 2007)
CG31872	Lipase	Association	Refractoriness	(Fiumera, Dumont, and Clark 2005)

[1]Association studies have found a correlation between Acp26Aa genotype and sperm competition parameters, but Acp26Aa null mutants do not appear to have any deficiencies in sperm competition assays (P1: Herndon and Wolfner 1999; P2: Christopher, Wong, and Wolfner unpublished data)

[2]Both knockout and association studies find a role for Acp36DE (Clark et al. 1995; Chapman et al. 2000), Acp29AB (Clark et al. 1995; Fiumera, Dumont, and Clark 2005), and Acp62F (Fiumera, Dumont, and Clark 2007; Mueller et al. 2008) in sperm competition.

2000; Fiumera, Dumont, and Clark 2005; Wong et al. in preparation). In addition, the sex-peptide and the prohormone ovulin contribute to egg-production and/or ovulation (Chen et al. 1988; Aigaki et al. 1991; Herndon and Wolfner 1995; Chapman et al. 2003; Liu and Kubli 2003; Heifetz et al. 2005). Such functional studies have ascribed roles to ~23 out of >100 known or suspected Acps (Table 1.1; Ravi Ram and Wolfner 2007), and as such much work remains to be done.

Studies in various insect species have also implicated non-protein components of the seminal fluid in post-mating responses. In some species of cricket, for example, prostaglandins transferred to the female in the male ejaculate induce female refactoriness to remating (Stanley-Samuelson and Loher 1986). In *Drosophila melanogaster*, a small hydrocarbon produced by the ejaculatory bulb, cis-vaccinyl acetate (CVA), is transferred to the female during copulation and exerts an anti-aphrodisiac effect on other males (Jallon, Antony, and Benamar 1981).

There is mounting evidence that some Acps are involved in immune defense. Lung and Wolfner (2001) demonstrated the presence of products with anti-bacterial activity in accessory gland extracts, and a recent study identified several specific Acps that reduce bacterial load upon systemic expression (Mueller, Page, and Wolfner 2007). Moreover, the anti-microbial peptide andropin (Samakovlis et al. 1991) is expressed only in the ejaculatory duct (which also contributes products to the male's seminal fluid), which suggests a specific role for this peptide following mating. In addition, mating has been shown to alter the expression levels of a number of immune genes (Lawniczak and Begun 2004; McGraw et al. 2004; Peng, Zipperlen, and Kubli 2005; Mack et al. 2006), with Acps playing an important role (McGraw et al. 2004; Peng, Zipperlen, and Kubli 2005). Interestingly, a recent study attributed systemic immune function to two genes whose expression in uninfected males is highly accessory-gland biased. Simultaneous RNAi knockdown of *sphinx1* and *sphinx2*,

putative protease homologs (i.e. proteins predicted to resemble proteases but bearing mutations likely to render them catalytically inactive), in the fat body and hemocytes increases the susceptibility of flies to gram-positive bacteria and fungi, apparently due to signaling roles in the Toll pathway (Kambris et al. 2006). In uninfected flies, expression of both genes is highly accessory gland biased in comparison to whole flies (75-fold and 133-fold, respectively; FlyAtlas.org). Barring off-target effects of the RNAi, low levels of *sphinx1* and *sphinx2* in the fat body and/or hemocytes may be sufficient for function, or expression could be induced following infection. The functions of these genes in the accessory gland, however, remain mysterious.

The evolution of Acps in Drosophila

The first evidence that *Drosophila* reproductive tract proteins may evolve more rapidly on average than do other classes of protein came from 2-D gel electrophoretic studies of male and female reproductive tracts (Civetta and Singh 1995). Civetta and Singh (1995) found that gonadal proteins, and particularly testis proteins, diverge much more rapidly between species than do non-gonadal proteins, on the basis of presence or absence of protein spots on 2-D gels. A number of subsequent studies have shown that Acps in particular tend to evolve rapidly at the amino acid level (e.g., (Mueller et al. 2005; Wagstaff and Begun 2005a); one recent study found that, for a group of 25 Acps, the rate of non-synonymous nucleotide substitution across six species of Drosophila is approximately twice the genome-wide average (Haerty et al. 2007).

There is ample evidence that the rapid protein evolution of some Acps is driven at least in part by positive selection, rather than merely by relaxed constraint (Table 1.2). Molecular population genetic analyses have documented evidence for positive selection on a number of Acp encoding loci (see Table 1.2), and comparisons

between species suggest that Acp encoding loci are more likely than genes associated with other functions to experience repeated episodes of positive selection (Haerty et al. 2007). Acps showing evidence for positive selection have a variety of known functions (Table 1.2), including the induction of ovulation (ovulin), sperm storage and competition (Acp29AB, Acp36DE), and the regulation of immune processes (*sphinx1* and *sphinx2*).

The repertoire of Acps in different species of *Drosophila* also appears to change rapidly. Attempts to find homologs of *D. melanogaster* Acps in other Drosophilids using reciprocal BLAST and/or syntenic considerations suggest that an unusual number of Acp loci are restricted to specific phylogenetic lineages (Figure 1.1; Mueller et al. 2005; Wagstaff and Begun 2005a; Haerty et al. 2007). Similarly, EST screens of the accessory glands of other Drosophilids (Begun and Lindfors 2005; Wagstaff and Begun 2005b; Begun et al. 2006) have identified numerous transcripts for putatively secreted proteins that are apparently absent from *D. melanogaster*. It is unclear, however, whether the apparently rapid turnover of Acp genes is indicative of differing selection pressures in different lineages, or of a relatively minor fitness cost associated with the loss of individual Acp genes (perhaps due to redundancy between Acps).

The rapid, adaptive evolution of reproductive proteins has been documented in a wide variety of species. Primate seminal fluid proteins (Kingan, Tatar, and Rand 2003; Clark and Swanson 2005), cricket Acps (Andres et al. 2006), gamete recognition proteins in plants (Chookajorn et al. 2004) and free-spawning marine invertebrates (Swanson, Aquadro, and Vacquier 2001; Galindo, Vacquier, and

Table 1.2 Evidence for positive selection on *D. melanogaster* Acp loci. Inferences of selection on putative Acp loci from *D. pseudoobscura* (Stevison, Counterman, and Noor 2004; Wagstaff and Begun 2005a; Schully and Hellberg 2006) or the desert Drosophilids (Wagstaff and Begun 2005b; Wagstaff and Begun 2007) not included.

Gene	Phenotypes	Evidence for selection	References
Sex-peptide	Egg-laying; refractoriness; toxicity; feeding	Intralocus linkage disequilibrium1	(Cirera and Aguadé 1997)
CG9997	Egg-laying; refractoriness; sperm release	MK	(Wong et al. 2008); Ch. 4 and 5
Ovulin	Ovulation; P1, P21	MK	(Aguadé, Miyashita, and Langley 1992; Aguadé 1998; Begun et al. 2000; Kern, Jones, and Begun 2004); (Tsaur and Wu 1997; Tsaur, Ting, and Wu 1998) Ch. 5
Acp36DE	Sperm storage; P1; P2	MK	(Begun et al. 2000); Ch. 5
Acp29AB	Sperm storage; P1; P2	MK	(Aguadé 1999; Begun et al. 2000; Holloway and Begun 2004)
Acp62F	P1, P22; fecundity; toxicity; ovulin processing	MK	Ch. 5
CG32382	Systemic immunity	MK	Ch. 5
CG32383	Systemic immunity	MK	Ch. 5
CG6069	?	HKA, PAML, MK	(Wong et al. 2008); Ch. 4, Ch. 5
CG32203	?	MK, PAML	(Haerty et al. 2007), Ch. 5
CG8137	?	MK	Ch. 5
CG33121	?	MK	Ch. 5
CG17242	?	MK	Ch. 5
CG10586	?	MK	Ch. 5
CG11664	?	MK	Ch. 5
CG10956	?	MK	Ch. 5
Acp32CD	?	PAML	(Haerty et al. 2007)
CG4847	?	PAML	(Haerty et al. 2007)
Pdi	?	PAML	(Haerty et al. 2007)
Lectin30A	?	MK	(Holloway and Begun 2004)
Acp53C14b, c	?	MK	(Holloway and Begun 2004)
Acp76A	?	MK	(Kern, Jones, and Begun 2004)

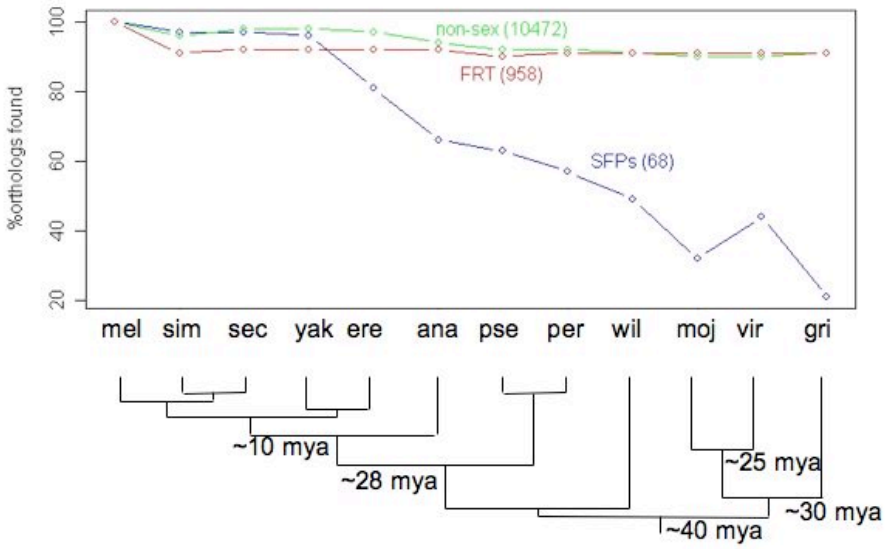


Figure 1.1 Evidence for lineage specific gain/loss of Acps in the genus Drosophila. Orthologs of *D. melanogaster* seminal fluid proteins ("SFPs"; largely consisting of Acps), female reproductive tract proteins ("FRTs") and non-reproductive tract proteins ("non-sex") were identified as reciprocal best blast hits in the genomes of 12 species of Drosophila. Adapted from Haerty et al. (2007).

Swanson 2003), and egg surface proteins in a broad range of species (e.g., Swanson et al. 2001) all show evidence of positive selection. In all but a few cases, however, the causes of such rapid evolution are unclear.

Potential causes of rapid reproductive protein evolution

Several hypotheses have been proposed to explain the unusual molecular evolution of reproductive proteins. As noted earlier, explanations invoking sexual selection (post-copulatory, in the case of internally fertilizing animals), are currently popular. Other hypotheses have been explored in less depth, but, as I hope will become clear, warrant consideration. Here, I outline two major hypotheses, and discuss empirical predictions and evidence cited in support of each.

(a) Post-copulatory sexual selection

"The war between the sexes is the only one in which both sides regularly sleep with the enemy"

- Quentin Crisp

Post-copulatory sexual selection is frequently proposed as a potential cause of rapid reproductive protein evolution. Under such scenarios, certain male genotypes are more fit when mated to certain female genotypes, and vice versa. Different varieties of post-copulatory sexual selection, e.g., direct benefit vs. sexual conflict, can result in different ensuing evolutionary dynamics. Three broad classes of sexual selection bear mentioning here: Direct benefit models, indirect benefit models, and sexual conflict models.

Under direct benefit models, a female receives direct fitness benefits from mating, for example in the form of food (e.g., Thornhill 1976) or nutrients in the

ejaculate (e.g., Lewis, Cratsley, and Rooney 2004). In *Photinus* fireflies, for example, males transfer a coiled, gelatinous spermatophore to females during mating. Sperm at the anterior end of the spermatophore enter the sperm storage organs, while the rest of the structure enters a specialized spermatophore-digesting gland. The receipt of multiple spermatophores increases a female's fecundity, probably as a result of egg provisioning - radiotracer studies demonstrate that proteins or amino acids from the spermatophore ultimately end up in the female's eggs, but not in her somatic tissues (reviewed in Lewis, Cratsley, and Rooney 2004). Similarly, radiolabelled amino acids in the ejaculate of male *Drosophila mojavensis*, but not of male *D. melanogaster*, are incorporated into both the soma and eggs of their mates (Markow and Ankney 1984). Phosphorous is also transferred to females in the male ejaculate of several Drosophila species, including *D. melanogaster* (Markow, Coppola, and Watts 2001).

In indirect benefit models, a female does not choose mates (or gametes) on the basis of resources imparted or offered to her. Rather, she accrues genetic benefits, for example in the form of offspring that are more viable or more attractive. Under the Fisherian 'sexy sons' hypothesis, one kind of indirect benefit model, the sole benefit that needs to be obtained by females from mating with an attractive male is the production of attractive sons (Kirkpatrick 1982). Importantly, if both female preference and male attractiveness are heritable, mate choice will generate linkage disequilibrium between the male trait locus (loci) and the female preference locus (loci). As a result, both male trait and female preference alleles can spread quickly through a population.

Sexual conflict models (Parker 1979; Arnqvist and Rowe 2005; Parker 2006) recognize that the interests of the two sexes may differ, potentially resulting in a struggle for control over key reproductive decisions. Not all sexual conflict takes place in the context of sexual selection (e.g., Rice 1984), but some conflict is expected over

certain parameters, e.g., female remating rate. All else being equal, it is in a male's best interest to prevent his partner from remating during a given mating cycle, such that his sperm fertilize all of his partner's eggs (if he is the first to mate). A female may, however, gain from multiple matings, perhaps by receiving nuptial gifts from several males, or by subsequent mating with a higher quality male. Under these conditions, males are expected to evolve adaptations to prevent female remating, and females are expected to evolve counter-adaptations. Specific predictions regarding the evolution of phenotypic traits or genetic loci engaged in sexually antagonistic coevolution vary widely between models (e.g., Gavrilets 2000; Gavrilets and Hayashi 2006; Parker 2006; Cameron, Day, and Rowe 2007), depending on the type of model used and its assumptions. Nonetheless, the rapid evolution of characters involved in sexually antagonistic coevolution is a prediction of some models (e.g., Holland and Rice 1998)). Interestingly, in *D. melanogaster*, several Acps are required to reduce a female's remating propensity (Chen et al. 1988; Aigaki et al. 1991; Chapman et al. 2003; Liu and Kubli 2003; Ram and Wolfner 2007), and at least one of these shows evidence for positive selection (CG9997 - Wong et al. 2008; Chapters 4 and 5).

Direct benefit, indirect benefit, and sexual conflict models predict, at least sometimes, positive selection on male and female traits. But, as we shall see below, this is not a unique prediction of post-copulatory sexual selection models, and so the observation of positive selection at Acp-encoding loci is not itself strong evidence in support of sexual selection. It is perhaps telling that there is now evidence for selection on a number of Acp genes whose products have known roles in reproductive processes (Table 1.2), but functional data are currently available only for a limited number of loci, and there may be some ascertainment bias (i.e., preferential analysis of genes with known reproductive phenotypes). Positive selection has also been documented on a number of loci encoding genes expressed in the female reproductive tract (Swanson

et al. 2004; Panhuis and Swanson 2006; Haerty et al. 2007; Lawniczak and Begun 2007), but again, functional characterization of these genes has only just begun.

Post-copulatory sexual selection models are supported by evidence for co-evolution between male and female traits or loci. Under any model of post-copulatory sexual selection that invokes female choice for male traits/molecules, male traits should 'track' female preference over evolutionary time. At the level of morphology, it has been shown in a number of insect taxa that sperm length correlates with the length of the female sperm storage organs and/or ducts, suggesting co-evolution between these two features (see Pitnick, Wolfner, and Suarez In press).

Comparative genomic and molecular population genetic analyses can also reveal evidence for co-evolution between male- and female-derived molecules, especially given detailed functional information regarding specific interactors. The strongest such example comes from the study of the abalone proteins lysin and VERL (Vitelline Envelope Receptor of Lysin) (reviewed in Panhuis, Clark, and Swanson 2006). Lysin is responsible for penetrating the vitelline envelope (VE) of the egg, and it is thought to do so by unraveling multimeric VERL molecules (Kresge, Vacquier, and Stout 2001). There is a substantial degree of species specificity in this interaction, as lysin proteins from different species of abalone show greater efficiency in dissolving the VE of homospecific eggs (Vacquier and Lee 1993).

Several aspects of lysin/VERL molecular evolution are suggestive of co-evolution. First, both molecules show strong evidence for positive selection, which, in light of their known interaction, is highly suggestive of co-evolution. Second, within a population of abalone, VERL is highly polymorphic, showing two distinct haplotype classes (Swanson, Aquadro, and Vacquier 2001); this pattern is predicted under at least one model of sexual conflict (Gavrilets and Waxman 2002), whereby females diverge into distinct "targets", leaving males "trapped" at a phenotypic middle ground.

Finally, polymorphism data are suggestive of linkage disequilibrium (LD) between lysin and VERL (Clark, Springer, and Swanson, in preparation). Such non-random associations are expected whenever certain female genotypes favor particular male genotypes, as in indirect benefit and sexual conflict models of sexual selection (e.g., Kirkpatrick 1982).

Data in flies specifically addressing co-evolution are virtually absent. Female receptors for rapidly evolving Acps are not currently known, and as such it is not possible to conduct tests of the kind described for lysin and VERL. However, once receptors have been identified, it will be instructive to look for patterns of co-evolution, including elevated LD within populations and correlated patterns of substitution between species (e.g., Dimmic et al. 2005). The first Acp receptor – SPR, the sex-peptide receptor – was recently reported (Yapici et al. 2008). Sex-peptide (SP) is a major contributor to several post-mating responses, including the induction of egg-laying and female resistance to re-mating, (Chen et al. 1988; Aigaki et al. 1991; Chapman et al. 2003; Liu and Kubli 2003), induction of feeding (Carvalho et al. 2006), and reduced female lifespan (Wigby and Chapman 2005). SP would thus seem to be a prime candidate for involvement in post-copulatory sexual selection (particularly sexual conflict). The evidence for positive selection on SP is ambiguous: High intralocus LD in a population sample from Spain (Cirera and Aguadé 1997) is suggestive of a non-neutral history, but demographic factors could be at play. In our study (Chapter 6), SP exhibits no evidence for positive selection. However, at least one molecule thought to modulate SP activity, CG9997 (Ravi Ram and Wolfner, in preparation), does show evidence for positive selection, as noted above. The SP-SPR interaction may then turn out to be an indirect focus of post-copulatory sexual selection via modulatory molecules.

Comparisons of rates of evolution between closely related taxa with different mating systems may also provide evidence that post-copulatory sexual selection operates on reproductive tract proteins. At the phenotypic level, it is well documented that male testis size correlates with female mating frequency in primates, supporting the hypothesis that increased sperm competition intensity leads to increased sperm production (Harcourt et al. 1981). Data from primates also suggest that selection on seminal fluid proteins is more intense in highly polyandrous species. Kingan, Tatar, and Rand (2003) and Jensen-Seaman and Li (2003) examined the evolution of semenogelins - semen coagulating proteins - in three primate taxa along a continuum of mating systems: chimpanzees (highly polyandrous), humans (moderately polyandrous), and gorillas (harem mating – largely monandrous). Semenogelins contribute to the primate mating plug, presumed to be an adaptation against subsequent sperm. Confirming predictions that the strength of selection should correlate with the risk of sperm competition, Kingan et al. (2003) found evidence for a recent selective sweep on semenogelin I in chimpanzees (they did not examine semenogelin II), while both studies found segregating stop codons in gorillas, suggestive of relaxed constraint.

Wagstaff and Begun (2005, 2007) and Kelleher et al. (2007) have performed similar comparisons of reproductive tract gene evolution between mating systems using distant relatives of *D. melanogaster*, the desert species *D. mojavensis* and *D. arizonae*. The latter two species remate much more frequently than do *D. melanogaster* and its close relatives (Markow 1982; Markow 1996), and comparative studies in the genus *Drosophila* suggest that species with high remating rates are more likely to have exaggerated ejaculates (Markow 2002). Consistent with these results, Wagstaff and Begun (2005) show increased rates of amino acid divergence and positive selection on putative Acp genes and testis genes in the desert species in

comparison to *D. melanogaster* and its relatives. Moreover, elevated numbers of young gene duplicates in the desert species suggest that gene family expansion is also influenced by sperm competition risk (Wagstaff and Begun 2005b). Similarly, Kelleher et al. (2007) found evidence for increased rates of adaptive evolution amongst female reproductive genes in the desert species, in comparison to *D. simulans*. Analyses of this kind show substantial promise in elucidating the origins of rapid reproductive gene evolution. Data from more species would be of use, since inferences from two sets of species are insufficient to make broad generalizations. It would be particularly instructive to examine Acp and female reproductive tract gene evolution in a monandrous species, such as *D. subobscura*; the clear prediction from post-copulatory sexual selection models is that Acps should show reduced levels of positive selection, and relaxed constraint, in such a species.

(b) Immunity

Host-pathogen interactions are thought to drive the rapid evolution of immune proteins in a wide variety of species (e.g., Hughes and Nei 1988; Sackton et al. 2007). For example, one recent study (Sackton et al. 2007) found that, in flies, genes whose products are involved in pathogen recognition are particularly prone to positive selection, presumably due to their direct interactions with rapidly evolving pathogen targets. It is thus tempting to posit that host-pathogen interactions in the female reproductive tract could result in positive selection on some proteins present in the female tract or in the seminal fluid (see Lawniczak et al. 2007 for a review). The risks of infection during mating are not clear for insects (Knell and Webberley 2004), although it is worth noting that the virulence of pathogens is generally expected to increase at higher transmission rates (Anderson and May 1982), such that higher rates

of remating may result in increased virulence of sexually transmitted diseases (Hamilton, 1990; Møller 1998).

Several pieces of evidence suggest an important role for immune processes after mating. As noted previously, the female reproductive tract and seminal fluid of flies includes a number of components with anti-microbial activity, as well as several proteases or protease inhibitors with immune modulatory activity. Several of these proteolysis regulators show evidence for positive selection (Chapter 6). In addition, microarray studies have shown that mating (and Acps specifically) changes the expression levels of a number of anti-microbial peptides in females (Lawniczak and Begun 2004; McGraw et al. 2004; Domanitskaya et al. 2007). The physiological consequences of mating for immunity, i.e., whether mating helps, hinders, or is irrelevant to, the female immune response, are currently unclear, however, owing to mixed results from different experimental techniques (Fedorka et al. 2007; Wigby et al. 2008).

The precise nature of the role played by immune molecules post-mating is not clear, and may or may not involve post-copulatory sexual selection. Immune modulators and AMPs expressed in the female reproductive tract could function solely to defend against pathogens, and undergo positive selection solely in response to this challenge. Alternatively (or concomitantly), immune molecules in the female reproductive tract may also create a hostile reproductive tract environment for sperm. In insects and mammals, the female reproductive tract can be an inhospitable place for sperm (e.g., Greeff and Parker 2000; Suarez and Pacey 2006; Holman and Snook 2008), with female-induced sperm mortality prior to fertilization. Sperm mortality may be a byproduct of immune processes, or may represent a method whereby sperm are selected in post-copulatory sexual selection (see Holman, Freckleton, and Snook

2008; Holman and Snook 2008; Pitnick, Wolfner, and Suarez In press) for more detailed discussion).

Correspondingly, immune molecules in the male seminal fluid may function to protect sperm from a hostile female reproductive tract (at least for immune modulators such as Sphinx), or again as a defense against pathogens. In the latter case, co-evolution with pathogens, rather than with females, is the most likely proximal cause of positive selection. In this case, then, natural selection rather than sexual selection would underlie the rapid evolution of some seminal fluid proteins.

Nonetheless, it should be noted that ejaculate components subject to natural selection due to host-pathogen interactions may also enter into post-copulatory sexual selection. Males whose ejaculate components enable their sperm to resist damage from pathogens, or from the female reproductive tract, may for example have improved sperm competitive ability due to sperm quality and/or quantity. Moreover, if immune molecules in the seminal fluid improve a female's chances of survival or her fecundity, then they may act as a "nuptial gift" in a direct benefits mode of sexual selection (Lawniczak et al. 2007). I would argue, however, that such consequences of immune function are not expected to generate the patterns predicted under male-female co-evolution (particularly correlated patterns of substitution and increased LD), since the protein's sequence evolves in response to interactions with pathogens rather than with the female.

Summary and future directions

Over the past ~15 years, researchers have accumulated a large body of evidence demonstrating that reproductive tract proteins – particularly those from males – tend to diverge rapidly between species, under the influence of positive selection. While it is generally thought that post-copulatory sexual selection underlies

this rapid divergence, rigorous tests of this hypothesis have rarely been conducted. The most convincing demonstrations come from statistical and functional evidence for male-female co-evolution in abalone, and from comparisons of rates of molecular evolution in species with different mating systems. In order to make broader generalizations, such approaches should be expanded in scope and phylogenetic distribution. In Drosophila in particular, identification of receptors for rapidly evolving Acps will facilitate statistical tests of co-evolution. Furthermore, as suggested above, the examination of reproductive tract protein evolution in species with a broader range of mating systems will allow more rigorous tests of the influence of polyandry.

Levels of control

I further suggest that a more comprehensive approach will be important in understanding female contributions to male-female co-evolution. Thus far, researchers have focused on co-evolution between directly interacting partners, e.g., lysin and VERL. However, in evolutionary terms, the important outcome is the reproductive phenotype – for example, female remating rate. While this outcome may be altered at the level of direct male-female interactions, other levels of control are possible. To continue the example, females could control their remating rate through sequence changes of the sex-peptide receptor that affect SP binding (direct interactions), or by modulating downstream events such as signal transduction and transcriptional activation. Alterations in the sequence or expression level of downstream effector genes could also play important roles. Thus, I submit that we have thus far only been looking at one aspect of a potentially very complicated system.

Approaches to identifying candidate female reproductive genes in *Drosophila* have largely focused on the female reproductive tract, since many Acps and sperm

bound proteins likely have their molecular partners in this organ. If my conjecture is right, however, we cannot expect all of the relevant molecules to be present at the site of insemination. Rather, neurological and endocrine pathways, as well as systems involved in resource partitioning, are likely to play important roles. As such, unbiased screens will be required to identify the full range of genes involved in the control of female reproductive processes. Yapici et al. 2008 conducted one such screen using RNA interference; this screen led to the identification of SPR. Since, however, some downstream effectors are likely to be shared amongst a variety of biological processes (e.g., signal transduction pathway components), systemic knockdown/knockout approaches will be insufficient due to pleiotropy. Instead, tissue specific and/or temporally controlled gene disruptions, traditional mutagenesis screens (which can generate hypomorphs, as well as regulatory and domain specific mutations), and QTL (e.g., Lawniczak and Begun 2005) and association studies (both of which make use of natural, non-lethal variation) will be useful in elucidating the mechanisms by which females control the post-mating response.

It is unclear *a priori* whether the same pathways, and the same components thereof, will contribute to changes in reproductive phenotypes in different lineages. Indeed, the rapid turnover of male Acp genes may indicate that different pathways or pathway components are the focus of post-mating male-female interactions in different species. According to this suggestion, males can manipulate their mates' responses, and/or females can distinguish between different potential fathers, using a variety of different mechanisms. If (under a conflict scenario) females evolve an effective mechanism to avoid male manipulation via one pathway, then male ejaculate proteins targeting that pathway will become ineffective, possibly leading to pseudogenization. Novel ejaculate proteins targeting other pathways (or other

components of the same pathway) might then become favored, leading to an accumulation of different ejaculate components in different lineages.

If variation at multiple levels of control does in fact contribute to the evolution of female post-mating responses, then expectations regarding patterns of molecular evolution under male-female co-evolution are less clear. First, we cannot expect that selection will operate preferentially on extra-cellular proteins, as has been frequently assumed (e.g., Swanson et al. 2004). Rather, we might predict that proteins (as well as regulatory sequences) at all levels of the relevant pathways, including intracellular signaling molecules and transcriptional activators, might be subject to positive selection. Second, correlated patterns of substitution between male- and female-reproductive proteins can only be expected for direct interactors, such as a receptor and its ligand. For a pathway with multiple inputs (e.g., several Acp ligands for several female receptors) and multiple intermediate steps influencing phenotypic outcomes, a model of one-to-one co-evolution may well be overly simplified. Explicit modeling of co-evolution in such a system may help to clarify the expected evolutionary dynamics.

The role of immune interactions in Acp evolution

I have suggested that host-pathogen interactions may also play important roles in driving the rapid evolution of some Acp genes. This hypothesis is currently difficult to test in a direct manner, in large part because virtually nothing is known about sexually transmitted diseases in any insect, including Drosophila (e.g., Lawniczak et al. 2007). If natural sexually transmitted pathogens are identified, it will be instructive to determine the consequences of specific Acp knockdown/knockouts, and of natural Acp variants, on pathogen transmission and proliferation. The roles of *sphinx1* and *sphinx2* will be particularly interesting to investigate, since the proteins encoded by

these genes are promising candidates for the induction of immune gene expression in females.

Consequences of selection on Acp genes

Ultimately, in order to understand the forces underlying rapid Acp evolution, it will be necessary to investigate the functional consequences of positive selection on Acp genes. As a first step, transgenic technologies will be useful for the comparison of Acp sequence variants in an otherwise isogenic background. Such an approach would allow for the functional assessment of Acp alleles from different strains, populations, or species, or of alleles bearing point mutations (say, of sites inferred to be under positive selection). Further transgenic studies might specifically investigate male-female co-evolution by determining the phenotypic consequences of particular receptor/ligand combinations.

In such transgenic studies, it is important to use an isogenic background in order to decisively establish that the gene of interest, and not other background variation, is responsible for the observed phenotype. This requirement, however, introduces a conceptual problem for evolutionary inferences (Lewontin 1974; Jensen, Wong, and Aquadro 2007). Genetic variation is abundant in natural populations of most species (including *D. melanogaster*), such that epistatic interactions can drastically alter the effects of variation at a single locus. "Background" variation may increase, reduce, or even reverse the effects of a single mutation. The net fitness effect of a single mutation will depend on the frequency and strength of modifiers in a population. To the extent that epistasis plays an important role in natural populations, the effect of a single mutation in an isogenic background may not accurately reflect its effect averaged over the full range of possible genotypes. Some headway on this problem might be gained by assaying a transgenic construct in a range of different

backgrounds, although in practice this is currently a difficult proposition. In cases of within-species variation, e.g., balancing selection or the fixation of a variant between populations, a combination of transgenic and association/QTL mapping studies may provide the best way to assess the phenotypic and fitness consequences of selection.

Thesis chapters

In chapters 2 and 3, I describe work on the self-interaction of the egg-laying hormone ovulin. We have found that ovulin forms a multimer (probably a dimer), and that coiled-coil interactions and a novel YxxxY motif are necessary for self-interaction. Interestingly, residues important for self-interaction are highly conserved across species, even though the rest of the ovulin protein evolves very rapidly. We propose that ovulin's coiled-coil tertiary structure poses limited constraints on its primary sequence, such that ample variation is available to selection. Indeed, coiled-coil proteins in the Drosophila genome tend to have a higher rate of amino acid divergence than do other proteins, suggesting that coiled-coil domains in general impose fewer constraints than do other structural domains.

In chapters 4 and 5, I attempt to address the issue of male-female co-evolution by focusing on a set of potentially interacting reproductive tract proteins: proteolysis regulators and targets of proteolysis. Using within- and between- species comparisons, we and others have found ample evidence for selection on proteolysis regulators (and targets) in both the male and female reproductive tracts, consistent with (although not exclusive to) co-evolution. We have also found evidence for positive selection on several proteins with known immune function, suggesting a role for host-pathogen interactions in driving the evolution of at least some reproductive tract proteins. We also find an excess of LD between proteolysis regulator (or target) loci expressed in different sexes, as predicted under co-evolutionary scenarios.

Finally, in chapter 6, I describe work on ancestral lineage sorting in the genus Drosophila. A reliable phylogeny is a prerequisite to many evolutionary analyses, including between-species inferences of selection as described in chapter 4. We found evidence for phylogenetic incongruence at two important nodes in the tree for the genus Drosophila, and provide evidence that this incongruence is due to ancient lineage sorting events.

REFERENCES

Aguadé, M. 1999. Positive selection drives the evolution of the Acp29AB accessory gland protein in Drosophila. Genetics **152**:543-551.

Aguadé, M. 1998. Different forces drive the evolution of the Acp26Aa and Acp26Ab accessory gland genes in the *Drosophila melanogaster* species complex. Genetics **150**:1079-1089.

Aguadé, M., N. Miyashita, and C. H. Langley. 1992. Polymorphism and divergence in the Mst26A male accessory gland gene region in Drosophila. Genetics **132**:755-770.

Aigaki, T., I. Fleischmann, P. S. Chen, and E. Kubli. 1991. Ectopic expression of sex peptide alters reproductive behavior of female *D. melanogaster*. Neuron **7**:557-563.

Anderson, R. M., and R. M. May. 1982. Coevolution of hosts and parasites. Parasitology **85 (Pt 2)**:411-426.

Andersson, M. 1994. Sexual Selection. Princeton University Press, Princeton, NJ.

Andres, J. A., L. S. Maroja, S. M. Bogdanowicz, W. J. Swanson, and R. G. Harrison. 2006. Molecular evolution of seminal proteins in field crickets. Mol Biol Evol **23**:1574-1584.

Arnqvist, G., and L. Rowe. 2005. Sexual conflict. Princeton University Press, Princeton.

Begun, D. J., and H. A. Lindfors. 2005. Rapid evolution of genomic Acp complement in the melanogaster subgroup of Drosophila. Mol Biol Evol **22**:2010-2021.

Begun, D. J., H. A. Lindfors, M. E. Thompson, and A. K. Holloway. 2006. Recently evolved genes identified from *Drosophila yakuba* and *D. erecta* accessory gland expressed sequence tags. Genetics **172**:1675-1681.

Begun, D. J., P. Whitley, B. L. Todd, H. M. Waldrip-Dail, and A. G. Clark. 2000. Molecular population genetics of male accessory gland proteins in Drosophila. Genetics **156**:1879-1888.

Birkhead, T. R., and A. P. Møller. 1998. Sperm competition and sexual selection. Academic Press, San Diego, CA.

Bloch Qazi, M. C., and M. F. Wolfner. 2003. An early role for the *Drosophila melanogaster* male seminal protein Acp36DE in female sperm storage. J Exp Biol **206**:3521-3528.

Braswell, W. E., J. A. Andres, L. S. Maroja, R. G. Harrison, D. J. Howard, and W. J. Swanson. 2006. Identification and comparative analysis of accessory gland proteins in Orthoptera. Genome **49**:1069-1080.

Cameron, E., T. Day, and L. Rowe. 2007. Sperm competition and the evolution of ejaculate composition. Am Nat **169**:e158-e172.

Carvalho, G. B., P. Kapahi, D. J. Anderson, and S. Benzer. 2006. Allocrine modulation of feeding behavior by the Sex Peptide of Drosophila. Curr Biol **16**:692-696.

Chapman, T., J. Bangham, G. Vinti, B. Seifried, O. Lung, M. F. Wolfner, H. K. Smith, and L. Partridge. 2003. The sex peptide of *Drosophila melanogaster*: female post-mating responses analyzed by using RNA interference. Proc Natl Acad Sci U S A **100**:9923-9928.

Chapman, T., and S. J. Davies. 2004. Functions and analysis of the seminal fluid proteins of male *Drosophila melanogaster* fruit flies. Peptides **25**:1477-1490.

Chapman, T., L. F. Liddle, J. M. Kalb, M. F. Wolfner, and L. Partridge. 1995. Cost of mating in *Drosophila melanogaster* females is mediated by male accessory gland products. Nature **373**:241-244.

Chapman, T., D. M. Neubaum, M. F. Wolfner, and L. Partridge. 2000. The role of male accessory gland protein Acp36DE in sperm competition in *Drosophila melanogaster*. Proc Biol Sci **267**:1097-1105.

Chen, P. S., E. Stumm-Zollinger, T. Aigaki, J. Balmer, M. Bienz, and P. Bohlen. 1988. A male accessory gland peptide that regulates reproductive behavior of female *D. melanogaster*. Cell **54**:291-298.

Chookajorn, T., A. Kachroo, D. R. Ripoll, A. G. Clark, and J. B. Nasrallah. 2004. Specificity determinants and diversification of the Brassica self-incompatibility pollen ligand. Proc Natl Acad Sci U S A **101**:911-917.

Cirera, S., and M. Aguadé. 1997. Evolutionary history of the sex-peptide (Acp70A) gene region in *Drosophila melanogaster*. Genetics **147**:189-197.

Civetta, A., and R. S. Singh. 1995. High divergence of reproductive tract proteins and their association with postzygotic reproductive isolation in *Drosophila melanogaster* and *Drosophila virilis* group species. J Mol Evol **41**:1085-1095.

Clark, A. G., M. Aguadé, T. Prout, L. G. Harshman, and C. H. Langley. 1995. Variation in sperm displacement and its association with accessory gland protein loci in *Drosophila melanogaster*. Genetics **139**:189-201.

Clark, N. L., J. E. Aagaard, and W. J. Swanson. 2006. Evolution of reproductive proteins from animals and plants. Reproduction **131**:11-22.

Clark, N. L., and W. J. Swanson. 2005. Pervasive Adaptive Evolution in Primate Seminal Proteins. PLoS Genet **1**:e35.

Collins, A. M., Caperna T.J., Williams V., Garrett W.M., and E. J.D. 2006. Proteomic analyses of male contributions to honey bee sperm storage and mating. Insect Molecular Biology.

Darwin, C. 1871. The Descent of Man, and Selection in Relation to Sex. John Murray, London.

Davies, S. J., and T. Chapman. 2006. Identification of genes expressed in the accessory glands of male Mediterranean Fruit Flies (*Ceratitis capitata*). Insect Biochem Mol Biol **36**:846-856.

Dimmic, M. W., M. J. Hubisz, C. D. Bustamante, and R. Nielsen. 2005. Detecting coevolving amino acid sites using Bayesian mutational mapping. Bioinformatics **21 Suppl 1**:i126-135.

Domanitskaya, E. V., H. Liu, S. Chen, and E. Kubli. 2007. The hydroxyproline motif of male sex peptide elicits the innate immune response in Drosophila females. FEBS J **274**:5659-5668.

Dottorini, T., L. Nicolaides, H. Ranson, D. W. Rogers, A. Crisanti, and F. Catteruccia. 2007. A genome-wide analysis in *Anopheles gambiae* mosquitoes reveals 46 male accessory gland genes, possible modulators of female behavior. Proc Natl Acad Sci U S A **104**:16215-16220.

Eberhard, W. G. 1996. Female control: Sexual selection by cryptic female choice. Princeton University Press, Princeton, N. J.

Fedorka, K. M., J. E. Linder, W. Winterhalter, and D. Promislow. 2007. Post-mating disparity between potential and realized immune response in *Drosophila melanogaster*. Proc Biol Sci **274**:1211-1217.

Fiumera, A. C., B. L. Dumont, and A. G. Clark. 2005. Sperm competitive ability in *Drosophila melanogaster* associated with variation in male reproductive proteins. Genetics **169**:243-257.

Fiumera, A. C., B. L. Dumont, and A. G. Clark. 2006. Natural variation in male-induced 'cost-of-mating' and allele-specific association with male reproductive genes in *Drosophila melanogaster*. Philos Trans R Soc Lond B Biol Sci **361**:355-361.

Fiumera, A. C., B. L. Dumont, and A. G. Clark. 2007. Associations between sperm competition and natural variation in male reproductive genes on the third chromosome of *Drosophila melanogaster*. Genetics **176**:1245-1260.

Galindo, B. E., V. D. Vacquier, and W. J. Swanson. 2003. Positive selection in the egg receptor for abalone sperm lysin. Proc Natl Acad Sci U S A **100**:4639-4643.

Gavrilets, S. 2000. Rapid evolution of reproductive barriers driven by sexual conflict. Nature **403**:886-889.

Gavrilets, S., and T. I. Hayashi. 2006. The dynamics of two- and three-way sexual conflicts over mating. Philos Trans R Soc Lond B Biol Sci **361**:345-354.

Gavrilets, S., and D. Waxman. 2002. Sympatric speciation by sexual conflict. Proc Natl Acad Sci U S A **99**:10533-10538.

Gillott, C. 2003. Male accessory gland secretions: modulators of female reproductive physiology and behavior. Annu Rev Entomol **48**:163-184.

Greeff, J. M., and G. A. Parker. 2000. Spermicide by females: what should males do? Proc Biol Sci **267**:1759-1763.

Haerty, W., S. Jagadeeshan, R. J. Kulathinal, A. Wong, K. Ravi Ram, L. K. Sirot, L. Levesque, C. G. Artieri, M. F. Wolfner, A. Civetta, and R. S. Singh. 2007. Evolution in the fast lane: rapidly evolving sex-related genes in Drosophila. Genetics **177**:1321-1335.

Harcourt, A. H., P. H. Harvey, S. G. Larson, and R. V. Short. 1981. Testis weight, body weight and breeding system in primates. Nature **293**:55-57.

Heifetz, Y., L. N. Vandenberg, H. I. Cohn, and M. F. Wolfner. 2005. Two cleavage products of the Drosophila accessory gland protein ovulin can independently induce ovulation. Proc Natl Acad Sci U S A **102**:743-748.

Herndon, L. A., and M. F. Wolfner. 1995. A Drosophila seminal fluid protein, Acp26Aa, stimulates egg laying in females for 1 day after mating. Proc Natl Acad Sci U S A **92**:10114-10118.

Holland, B., and W. R. Rice. 1998. Perspective: Chase-away sexual selection: Antagonistic seduction versus resistance. Evolution **52**:1-7.

Holloway, A. K., and D. J. Begun. 2004. Molecular evolution and population genetics of duplicated accessory gland protein genes in Drosophila. Mol Biol Evol **21**:1625-1628.

Holman, L., R. P. Freckleton, and R. R. Snook. 2008. What use is an infertile sperm? A comparative study of sperm-heteromorphic Drosophila. Evolution Int J Org Evolution **62**:374-385.

Holman, L., and R. R. Snook. 2008. A sterile sperm caste protects brother fertile sperm from female-mediated death in *Drosophila pseudoobscura*. Curr Biol **18**:292-296.

Hughes, A. L., and M. Nei. 1988. Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. Nature **335**:167-170.

Jallon, J. M., C. Antony, and O. Benamar. 1981. Un antiaphrodisiaque produit par les mâles de *Drosophila melanogaster* et transféré aux femelles lors de la copulation. Comptes rendus de l'académie des sciences de Paris, Série III **292**:1147-1149.

Jensen, J. D., A. Wong, and C. F. Aquadro. 2007. Approaches for identifying targets of positive selection. Trends Genet **23**: 568-577.

Jensen-Seaman, M. I., and W. H. Li. 2003. Evolution of the hominoid semenogelin genes, the major proteins of the ejaculated semen. J Mol Evol **57**: 261-270.

Kalb, J. M., A. J. DiBenedetto, and M. F. Wolfner. 1993. Probing the function of *Drosophila melanogaster* accessory glands by directed cell ablation. Proc Natl Acad Sci U S A **90**:8093-8097.

Kambris, Z., S. Brun, I. H. Jang, H. J. Nam, Y. Romeo, K. Takahashi, W. J. Lee, R. Ueda, and B. Lemaitre. 2006. Drosophila immunity: a large-scale in vivo RNAi screen identifies five serine proteases required for Toll activation. Curr Biol **16**:808-813.

Kelleher, E. S., W. J. Swanson, and T. A. Markow. 2007. Gene duplication and adaptive evolution of digestive proteases in *Drosophila arizonae* female reproductive tracts. PLoS Genet **3**:1541-1549.

Kern, A. D., C. D. Jones, and D. J. Begun. 2004. Molecular population genetics of male accessory gland proteins in the *Drosophila simulans* complex. Genetics **167**:725-735.

Kingan, S. B., M. Tatar, and D. M. Rand. 2003. Reduced polymorphism in the chimpanzee semen coagulating protein, semenogelin I. J Mol Evol **57**:159-169.

Kirkpatrick, M. 1982. Sexual selection and the evolution of female choice. Evolution **36**:1-12.

Knell, R. J., and K. M. Webberley. 2004. Sexually transmitted diseases of insects: distribution, evolution, ecology and host behaviour. Biol Rev Camb Philos Soc **79**:557-581.

Kresge, N., V. D. Vacquier, and C. D. Stout. 2001. Abalone lysin: the dissolving and evolving sperm protein. Bioessays **23**:95-103.

Lawniczak, M. K., A. I. Barnes, J. R. Linklater, J. M. Boone, S. Wigby, and T. Chapman. 2007. Mating and immunity in invertebrates. Trends Ecol Evol **22**:48-55.

Lawniczak, M. K., and D. J. Begun. 2004. A genome-wide analysis of courting and mating responses in *Drosophila melanogaster* females. Genome **47**:900-910.

Lawniczk, M. K., and D. J. Begun. 2005. A QTL analysis of female variation contributing to refractoriness and sperm competition in *Drosophila melanogaster*. Genet Res **86**: 107-114.

Lawniczak, M. K., and D. J. Begun. 2007. Molecular Population Genetics of Female-expressed Mating-induced Serine Proteases in *Drosophila melanogaster*. Mol Biol Evol. **24**: 1944-1951.

Lewis, S. M., C. K. Cratsley, and J. A. Rooney. 2004. Nuptial gifts and sexual selection in Photnius fireflies. Integrative and Comparative Biology **44**: 234-237.

Lewontin, R. C. 1974. The genetic basis of evolutionary change. Columbia University Press.

Liu, H., and E. Kubli. 2003. Sex-peptide is the molecular basis of the sperm effect in *Drosophila melanogaster*. Proc Natl Acad Sci U S A **100**:9929-9933.

Lung, O., U. Tram, C. M. Finnerty, M. A. Eipper-Mains, J. M. Kalb, and M. F. Wolfner. 2002. The *Drosophila melanogaster* seminal fluid protein Acp62F is a protease inhibitor that is toxic upon ectopic expression. Genetics **160**:211-224.

Mack, P. D., A. Kapelnikov, Y. Heifetz, and M. Bender. 2006. Mating-responsive genes in reproductive tissues of female *Drosophila melanogaster*. Proc Natl Acad Sci U S A **103**:10358-10363.

Markow, T. A. 1996. Evolution of Drosophila mating systems. Evol Biol **29**:73-106.

Markow, T. A. 2002. Perspective: female remating, operational sex ratio, and the arena of sexual selection in Drosophila species. Evolution Int J Org Evolution **56**:1725-1734.

Markow, T. A. 1982. Mating systems of cactophilic Drosophila. Pp. 273-287 *in* J. S. F. Barker, and W. T. Starmer, eds. Ecological genetics and evolution: The cactus-yeast-Drosophila model system. Plenum Press, New York.

Markow, T. A., and P. F. Ankney. 1984. Drosophila Males Contribute to Oogenesis in a Multiple Mating Species. Science **224**:302-303.

Markow, T. A., A. Coppola, and T. D. Watts. 2001. How Drosophila males make eggs: it is elemental. Proc Biol Sci **268**:1527-1532.

McGraw, L. A., G. Gibson, A. G. Clark, and M. F. Wolfner. 2004. Genes regulated by mating, sperm, or seminal proteins in mated female *Drosophila melanogaster*. Curr Biol **14**:1509-1514.

Møller, A. P. 1998. Sperm competition and sexual selection *in* T. R. Birkhead, and A. P. Møller, eds. Sperm competition and sexual selection. Academic Press, San Diego.

Mueller, J. L., J. Linklater, K. Ravi Ram, T. Chapman, and M. F. Wolfner. 2008. Targeted gene deletion and phenotypic analysis of the *Drosophila melanogaster* seminal fluid protease inhibitor Acp62F. Genetics.

Mueller, J. L., J. L. Page, and M. F. Wolfner. 2007. An ectopic expression screen reveals the protective and toxic effects of Drosophila seminal fluid proteins. Genetics **175**:777-783.

Mueller, J. L., K. R. Ram, L. A. McGraw, M. C. Bloch Qazi, E. D. Siggia, A. G. Clark, C. F. Aquadro, and M. F. Wolfner. 2005. Cross-species comparison of Drosophila male accessory gland protein genes. Genetics **171**:131-143.

Neubaum, D. M., and M. F. Wolfner. 1999. Mated *Drosophila melanogaster* females require a seminal fluid protein, Acp36DE, to store sperm efficiently. Genetics **153**:845-857.

Panhuis, T. M., N. L. Clark, and W. J. Swanson. 2006. Rapid evolution of reproductive proteins in abalone and Drosophila. Philos Trans R Soc Lond B Biol Sci **361**:261-268.

Panhuis, T. M., and W. J. Swanson. 2006. Molecular evolution and population genetic analysis of candidate female reproductive genes in Drosophila. Genetics **173**:2039-2047.

Parker, G. A. 2006. Sexual conflict over mating and fertilization: an overview. Philos Trans R Soc Lond B Biol Sci **361**:235-259.

Parker, G. A. 1979. Sexual selection and sexual conflict. Pp. 123-166 *in* M. S. Blum, and N. A. Blum, eds. Sexual selection and reproductive competition in insects. Academic Press, London.

Peng, J., P. Zipperlen, and E. Kubli. 2005. Drosophila sex-peptide stimulates female innate immune system after mating via the Toll and Imd pathways. Curr Biol **15**:1690-1694.

Pitnick, S., M. F. Wolfner, and S. S. Suarez. In press. Ejaculate- and sperm-female interactions *in* T. R. Birkhead, D. J. Hosken, and S. Pitnick, eds. Sperm biology: An evolutionary perspective. Elsevier Press.

Ram, K. R., and M. F. Wolfner. 2007. Sustained Post-Mating Response in *Drosophila melanogaster* Requires Multiple Seminal Fluid Proteins. PLoS Genet **3**:e238.

Ravi Ram, K., L. K. Sirot, and M. F. Wolfner. 2006. Predicted seminal astacin-like protease is required for processing of reproductive proteins in *Drosophila melanogaster*. Proc Natl Acad Sci U S A **103**:18674-18679.

Ravi Ram, K., and M. F. Wolfner. 2007. Seminal influences: Drosophila Acps and the molecular interplay between males and females during reproduction. Integrative and Comparative Biology.

Rice, W. R. 1984. Sex chromosomes and the evolution of sexual dimorphism. Evolution **38**: 735-742.

Sackton, T. B., B. P. Lazzaro, T. A. Schlenke, J. D. Evans, D. Hultmark, and A. G. Clark. 2007. Dynamic evolution of the innate immune system in Drosophila. Nat Genet **39**:1461-1468.

Samakovlis, C., P. Kylsten, D. A. Kimbrell, A. Engstrom, and D. Hultmark. 1991. The andropin gene and its product, a male-specific antibacterial peptide in *Drosophila melanogaster*. Embo J **10**:163-169.

Schully, S. D., and M. E. Hellberg. 2006. Positive Selection on Nucleotide Substitutions and Indels in Accessory Gland Proteins of the *Drosophila pseudoobscura* Subgroup. J Mol Evol.

Sirot, L. K., R. L. Poulson, M. C. McKenna, H. Girnary, M. F. Wolfner, and L. C. Harrington. 2008. Identity and transfer of male reproductive gland proteins of the dengue vector mosquito, *Aedes aegypti*: potential tools for control of female feeding and reproduction. Insect Biochem Mol Biol **38**:176-189.

Stanley-Samuelson, D. W., and W. Loher. 1986. Prostaglandins in insect reproduction. Ann Entomol Soc Am **79**:841-853.

Stevison, L. S., B. A. Counterman, and M. A. Noor. 2004. Molecular evolution of X-linked accessory gland proteins in *Drosophila pseudoobscura*. J Hered **95**:114-118.

Suarez, S. S., and A. A. Pacey. 2006. Sperm transport in the female reproductive tract. Hum Reprod Update **12**:23-37.

Swanson, W. J., C. F. Aquadro, and V. D. Vacquier. 2001. Polymorphism in abalone fertilization proteins is consistent with the neutral evolution of the egg's receptor for lysin (VERL) and positive darwinian selection of sperm lysin. Mol Biol Evol **18**:376-383.

Swanson, W. J., and V. D. Vacquier. 2002. The rapid evolution of reproductive proteins. Nat Rev Genet **3**:137-144.

Swanson, W. J., A. Wong, M. F. Wolfner, and C. F. Aquadro. 2004. Evolutionary expressed sequence tag analysis of Drosophila female reproductive tracts identifies genes subjected to positive selection. Genetics **168**:1457-1465.

Swanson, W. J., Z. Yang, M. F. Wolfner, and C. F. Aquadro. 2001. Positive Darwinian selection drives the evolution of several female reproductive proteins in mammals. Proc Natl Acad Sci U S A **98**:2509-2514.

Thornhill, R. 1976. Sexual selection and nuptial feeding behavior in *Bittacus apicalis* (Insecta: Mecoptera). Am Nat **110**:529-548.

Tram, U., and M. F. Wolfner. 1998. Seminal fluid regulation of female sexual attractiveness in *Drosophila melanogaster*. Proc Natl Acad Sci U S A **95**:4051-4054.

Tsaur, S. C., C. T. Ting, and C. I. Wu. 1998. Positive selection driving the evolution of a gene of male reproduction, Acp26Aa, of Drosophila: II. Divergence versus polymorphism. Mol Biol Evol **15**:1040-1046.

Tsaur, S. C., and C. I. Wu. 1997. Positive selection and the molecular evolution of a gene of male reproduction, Acp26Aa of Drosophila. Mol Biol Evol **14**:544-549.

Vacquier, V. D., and Y. H. Lee. 1993. Abalone sperm lysin: unusual mode of evolution of a gamete recognition protein. Zygote **1**:181-196.

Wagstaff, B. J., and D. J. Begun. 2005a. Comparative genomics of accessory gland protein genes in *Drosophila melanogaster* and *D. pseudoobscura*. Mol Biol Evol **22**:818-832.

Wagstaff, B. J., and D. J. Begun. 2005b. Molecular Population Genetics of Accessory Gland Protein Genes and Testis-expressed Genes in *Drosophila mojavensis* and *D. arizonae*. Genetics.

Wagstaff, B. J., and D. J. Begun. 2007. Adaptive evolution of recently duplicated accessory gland protein genes in desert Drosophila. Genetics **177**:1023-1030.

Wigby, S., and T. Chapman. 2005. Sex peptide causes mating costs in female *Drosophila melanogaster*. Curr Biol **15**:316-321.

Wigby, S., E. V. Domanitskaya, Y. Choffat, E. Kubli, and T. Chapman. 2008. The effect of mating on immunity can be masked by experimental piercing in female *Drosophila melanogaster*. J Insect Physiol **54**:414-420.

Wong, A., M. C. Turchin, M. F. Wolfner, and C. F. Aquadro. 2008. Evidence for positive selection on *Drosophila melanogaster* seminal fluid protease homologs. Mol Biol Evol **25**:497-506.

Xue, L., and M. Noll. 2000. Drosophila female sexual behavior induced by sterile males showing copulation complementation. Proc Natl Acad Sci U S A **97**:3272-3275.

Yapici, N., Y. J. Kim, C. Ribeiro, and B. J. Dickson. 2008. A receptor that mediates the post-mating switch in Drosophila reproductive behaviour. Nature **451**:33-37.

CHAPTER 2

EVIDENCE FOR STRUCTURAL CONSTRAINT ON OVULIN, A RAPIDLY EVOLVING *DROSOPHILA MELANOGASTER* SEMINAL PROTEIN[1]

Introduction

Adaptive pressures and functional constraint may exert opposing forces on the sequences and three dimensional structures of proteins. The possibility for evolutionary novelty is limited, sometimes severely, by structural and physico-chemical features required for protein function. Well documented cases of extreme conservation over large evolutionary distances are illustrative of the role of functional constraint (DeLange et al. 1969). By contrast, many genes involved in reproduction and immunity appear to have undergone repeated episodes of adaptive evolution, with high levels of sequence diversity within and between species (Hughes and Nei 1988; Swanson and Vacquier 2002).

For the proteins encoded by such rapidly evolving genes, the strength and targets of functional constraint may be less obvious. The available evidence suggests that some level of constraint is required for even the most diverse proteins to retain function. Indeed, the presence of a stable tertiary structure may reduce functional constraint on other portions of a protein (Bloom et al. 2006), reducing the tension between adaptive pressures and functional constraint. In the case of the highly polymorphic crucifer self-incompatibility ligand SCR, for example, primary sequence conservation is very limited - only a few key residues (primarily cysteines) are

[1] This chapter was previously published as: Wong A, Albright SN, Wolfner MF. 2006. Evidence for structural constraint on ovulin, a rapidly evolving *Drosophila melanogaster* seminal protein. *Proc Natl Acad Sci* USA 103:18644-9. Shannon Albright performed the yeast-two hybrid screen and subsequent yeast-two hybrid experiments using ovulin cleavage fragments. I performed all other experimental work, and co-wrote the manuscript with MFW. Copyright permissions for theses are automatically granted by the journal.

conserved between alleles. Structural modeling suggests, however, that overall secondary and tertiary structure is maintained (Chookajorn et al. 2004).

The *D. melanogaster* seminal fluid protein ovulin represents a particularly striking case of rapid protein evolution. Ovulin is a 264 amino acid polypeptide that is produced in the male accessory gland, along with about 50-100 other accessory gland proteins (Acps). Acps are transferred to the female along with sperm and other secretions during copulation, and are known to cause a variety of physiological and behavioral changes in females (for reviews, see Wolfner 2002; Bloch Qazi, Heifetz, and Wolfner 2003; Gillott 2003; Kubli 2003; Chapman and Davies 2004; Wolfner, Heifetz, and Applebaum 2005). After transfer to the female, ovulin is sequentially cleaved into four smaller peptides (CP1, CP2, CP3C, and CP3N; Park and Wolfner 1995; Heifetz et al. 2005), and increases ovulation in females during the first 24 hours after mating (Monsma and Wolfner 1988; Herndon and Wolfner 1995; Heifetz et al. 2000). Full length ovulin, as well as its two C-terminal cleavage products, are each individually capable of inducing ovulation in ectopic expression assays (Heifetz et al. 2005). Ovulin may therefore act as a prohormone, in that its cleavage may release active products; however, that uncleaved ovulin is also active contrasts with other prohormones (Derynck et al. 1985; Seidah et al. 1999). Remarkably, the most C-terminal cleavage product of ovulin contains a region of sequence similarity to a family of egg-laying hormones (the ELHs and Califins) from *Aplysia* (Monsma and Wolfner 1988; Heifetz et al. 2000).

Ovulin has evolved extremely rapidly at the amino acid level (Aguadé, Miyashita, and Langley 1992; Tsaur and Wu 1997; Aguadé 1998; Tsaur, Ting, and Wu 1998). Amino acid divergence between the closely related species *D. melanogaster* and *D. simulans* is about 15% for ovulin, while average divergence for other proteins is only about 1-2% (Tamura, Subramanian, and Kumar 2004;

Andolfatto 2005). Population genetic analyses in several species suggest that at least part of ovulin's divergence is driven by positive selection, whereby some new amino acid variants fix rapidly in a population owing to fitness benefits that they confer (Aguadé, Miyashita, and Langley 1992; Tsaur and Wu 1997; Aguadé 1998; Tsaur, Ting, and Wu 1998). Thus, it appears that ovulin variants can confer significant benefits to a male, presumably due to some advantage that they grant in the context of post-mating events, such as stimulation of egg production.

Although the precise nature of this advantage is not presently clear, ovulin's rapid evolution mirrors that of other reproductive traits. Evolutionary biologists have long noted that some traits involved in mating, e.g. sperm length (Pitnick, Markow, and Spicer 1995) and male genital morphology (Kopp and True 2002), diverge rapidly between species. Variation in such traits may contribute to reproductive success, for example by influencing an individual's mating opportunities or control over reproductive decisions. As a result, sexual selection can rapidly fix favorable variants in a population, leading to the rapid evolution of reproductive traits. At the molecular level, proteins involved in sperm competition (competition between sperm from different males within the reproductive tract of a single female) or in sexually antagonistic co-evolution may similarly experience strong sexual selection. Association studies have suggested that ovulin may influence sperm competition, most likely in the 'offense' component (Clark et al. 1995; Herndon and Wolfner 1995; Fiumera, Dumont, and Clark 2005), raising a potential cause for ovulin's rapid evolution. Alternatively, conflict between males and females over the rate of egg-laying may result in sexually antagonistic co-evolution between ovulin and its (as yet unidentified) receptor in females.

While ovulin's rapid evolution has been described in some detail, nothing is known about the structural constraints (if any) that are necessary for its function. In

order to fully understand ovulin's molecular evolution and function as a hormone (or prohormone), structure/function relationships within the ovulin polypeptide need to be clarified. To date, cleavage products of ovulin sufficient to induce ovulation have been identified (Heifetz et al. 2005). Further work will be required to determine the functional roles of rapidly evolving regions and amino acid residues, as well as those domains and residues that are more highly conserved. If some form of sexual selection does in fact operate on ovulin, then one might predict that rapidly evolving residues will be involved in interactions with proteins produced by the female, or in sperm competition between males. The rapid evolution of some sites, together with maintenance of ovulin function, may itself be made possible by an evolutionarily stable structural backbone.

Here, we continue investigation of the structure and function of ovulin using biochemical and computational methods. We report that ovulin interacts with itself in several assays, with a predicted coiled-coil in its C-terminus likely playing a major role. Residues predicted to be critical for self-interaction are conserved relative to the rest of the ovulin protein, consistent with our expectations. We propose that elements involved in self-interaction form a conserved structural backbone for the ovulin protein, resulting in greater evolutionary flexibility at other sites.

Materials and Methods

Yeast two-hybrid analysis

A yeast two-hybrid screen for interactors of ovulin was performed using the Matchmaker system, according to the manufacturer's protocol (Clontech). A mixed-sex adult cDNA library containing $3.5x10^6$ independent clones ('prey') was screened for interactors with full-length ovulin minus its predicted signal sequence ('bait'). Bait/prey interactions were detected by histidine prototrophy and X-gal staining. Full-

length sequences of the interacting genes were identified by BLAST against the complete *D. melanogaster* genome sequence (Adams et al. 2000).

Coding sequences for each of the putative ovulin cleavage products CP1, CP2, CP3N , and CP3C (cloned as described in Heifetz et al. 2005) were sub-cloned using the Gateway system (Invitrogen) into Gateway-compatible yeast-two hybrid vectors (original vectors from Clontech, modified by Ravi Ram K., A. Garfinkel and M. Wolfner, unpublished). All possible pairwise interactions were then tested, using histidine and adenine prototrophy, and X-gal staining, as markers.

Production and purification of GST fusion proteins

Secreted GST and GST-ovulin (minus ovulin's native signal sequence) were produced in 293T cells using pGST and pGST-GW, derivatives of the pAP-TAG5 vector (GenHunter). We were unable to obtain expression of GST-tagged CG13083 in 293T cells using pGST-GW. As an alternative, recombinant GST-CG13083 was produced in BL21-AI *E. coli* using pDEST-15 (Invitrogen).

Fusion proteins were purified on 50 µl of glutathione coated agarose beads (Sigma) overnight at 4˚C, using 1 ml of medium (GST and GST-ovulin) or 500 µl of bacterial lysate (GST-CG13083). Beads were washed 3 times with 1 ml phosphate buffered saline (PBS) + protease inhibitor cocktail (Roche), and stored in the same buffer (modified from Swaffield and Johnston 1996). Production and purification of a protein of the correct predicted molecular weight was verified by Coomassie staining of SDS-PAGE gels, and by Western blotting using antibodies to GST (Sigma) or ovulin.

GST-pulldown assays

GST-pulldown assays were performed using extracts from the accessory glands of 10 3-5 day post-eclosion Canton-S males, or from the reproductive tracts of ~20 mated 3-5 day old Canton-S females (30-90 minutes post-mating). Tissues of interest were dissected into 50µl 40% sucrose + protease inhibitor cocktail (Roche), and were homogenized using a plastic homogenizer. 1ml of NP-40 buffer (50mM Tris pH 7.5, 150mM NaCl, 0.5% NP-40, 10% glycerol, 3mM $MgCl_2$, 1mM EDTA) + protease inhibitor cocktail was then added, along with 10-50µl of glutathione-agarose beads with bound GST, GST-ovulin, or GST-CG13083. Following overnight incubation with rotation at 4˚C, the supernatant was removed and the beads were washed 5 times with NP-40 buffer + protease inhibitor cocktail. Following SDS-PAGE, ovulin was detected by Western blotting using anti-ovulin antibodies at a concentration of 1:2000.

Bioinformatic and evolutionary analyses

Signal sequence prediction was carried out using SignalP 3.0 (Bendtsen et al. 2004), and transmembrane domain predictions were performed using Sosui (Hirokawa, Boon-Chieng, and Mitaku 1998) and HMMTOP (Tusnady and Simon 2001). Prediction of secondary stucture was carried out using PsiPred (McGuffin, Bryson, and Jones 2000), and putative coiled-coil domains were identified using Coils (Lupas, Van Dyke, and Stock 1991) . For interspecific sequence comparisons, we obtained coding sequences of ovulin from a single individual each of *D. melanogaster* (GenBank accession no. NM_057296), *D. simulans* (strain sim1, GenBank accession no. AY499205), and *D. pseudoobscura* (strain pse1, GenBank accession no. AY818043). The number of nonsynonymous (amino acid changing) nucleotide substitutions per nonsynonymous site (K_a) between *D. melanogaster* and *D. simulans*

was calculated in windows of 30 nucleotides, every 10 nt along the coding sequence, using DnaSP4.1 (Rozas et al. 2003). This sliding window analysis permits visualization of amino acid divergence in different regions of the protein.

Non-reducing SDS-PAGE

Male accessory glands (2 individuals) or the reproductive tracts of mated females (45-90 minutes post-mating, 3 individuals) were dissected into 50µl 40% sucrose + protease inhibitor cocktail (Roche), and homogenized. His_6-ovulin was produced in *E. coli* using pDEST-17 (Invitrogen). Protein extracts were mixed with an equal volume of 2x SDS-PAGE loading buffer, with no reducing agent or with 0.1% β-ME. Samples were then subjected to 15% SDS-PAGE, and ovulin was detected by Western blotting using anti-ovulin antibodies at a concentration of 1:2000.

Results

Ovulin interacts with itself

In a yeast two-hybrid screen for proteins that can interact with full length ovulin, 61 individual interactors were identified. Sequencing indicated that they corresponded to 14 different genes (Table 2.1). Since ovulin is an extracellular protein, we reasoned that its molecular partner(s) would likely also be extracellular or on the cell surface. 3 of the 14 candidates encode proteins with predicted signal sequences, CG13083 and CG32642 (predicted ORFs - Adams et al. 2000), and ovulin itself. Three distinct ovulin clones were recovered in our two-hybrid screen, encoding amino acids 25-264, amino acids 128-264, or the C-terminal 45 amino acids of ovulin (amino acids 219-264). This final fragment will be referred to as ovulin C45.

Because yeast two-hybrid analysis can yield spurious interactions, we further tested whether the three candidate interactors could in fact interact with ovulin. First,

Table 2.1 Yeast two-hybrid interactors of ovulin. No. of hits, number of clones identified in screen that correspond to gene; SS, predicted signal sequence; TM, predicted transmembrane domain.

Gene name	No. of hits	SS/TM
CG8982 (Ovulin)	13	SS
CG13083	2	SS
CG32642	5	SS
CG3815	1	None
CG6392 (CENP-meta)	16	None
CG7773 (fidipidine)	2	None
CG31907	4	None
CG16747 (Oda)	6	None
CG13949	1	None
CG5934	2	None
CG3184 (Unc-76)	3	None
CG3981	1	None
CG9391	1	None
CG31175 (Dystrophin	4	None

we used RT-PCR to determine if the expression patterns of CG13083 and CG32642 are such that ovulin could normally encounter their protein products. We failed to confirm expression of CG32642 in adult *D. melanogaster*, suggesting that the yeast two-hybrid interaction is not meaningful in an *in vivo* context. CG13083 is expressed in adult females; males and larval stages were not assayed.

We next performed GST-pulldown assays between ovulin and the remaining two candidate interactors (ovulin and CG13083). We confirmed the self-interaction of ovulin (Figure 2.1, lanes 1-5). An N-terminal GST fusion of the full predicted secreted portion of ovulin produced in mammalian tissue culture cells was able to pull down native ovulin from extracts of male accessory glands. A similar quantity of GST alone, by contrast, was unable to pull down native ovulin. We were unable to confirm interaction between CG13083 and ovulin using GST-pulldown assays, using a GST-CG13083 fusion protein produced in *E. coli*. It should be noted that this result does not disconfirm an interaction between ovulin and CG13083, as the fusion protein may not be appropriately folded, or otherwise post-translationally modified. However, we do not pursue any further characterization of this potential interactor here.

The C-terminus of ovulin interacts with itself

Upon transfer to the female, ovulin is sequentially cleaved into a number of smaller fragments (Monsma, Harada, and Wolfner 1990; Park and Wolfner 1995). To further delineate the regions of ovulin involved in its self-interaction, we performed a yeast-two hybrid analysis with all pairwise combinations of ovulin's putative cleavage products (Figure 2.2). Only diploid yeast expressing both activation domain-CP3C and DNA binding domain-CP3C fusion proteins activated the *HIS3*, *ADE2*, and *lacZ* reporter genes (Figure 2.2a). Thus, ovulin's most C-terminal processing product CP3C interacts with itself in the two-hybrid system. CP3C did not interact with any

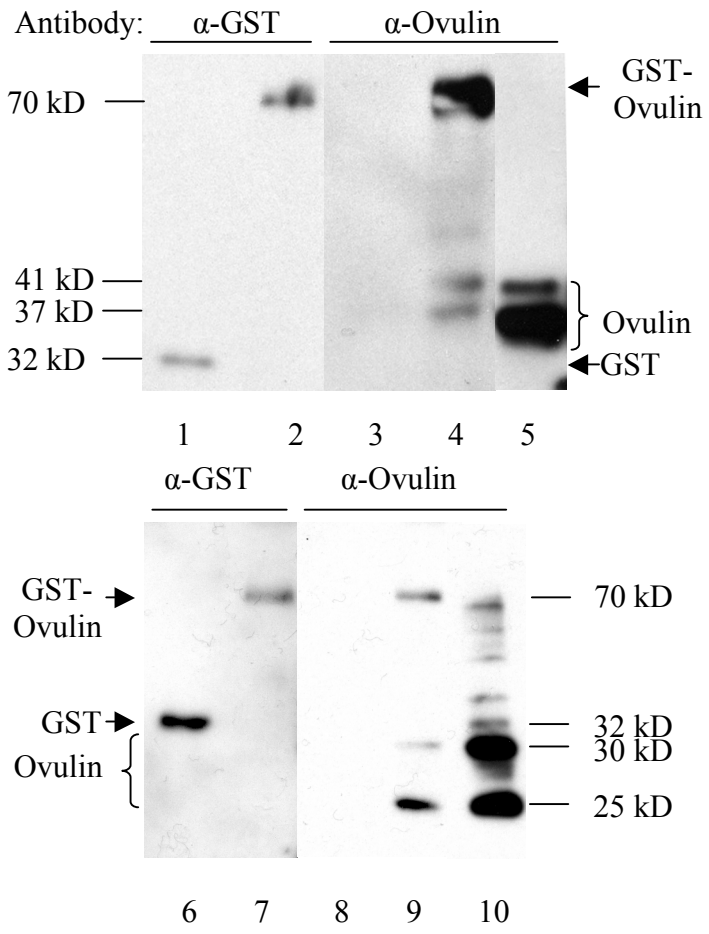
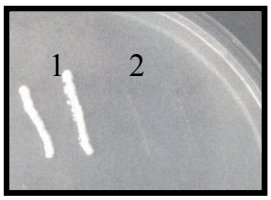
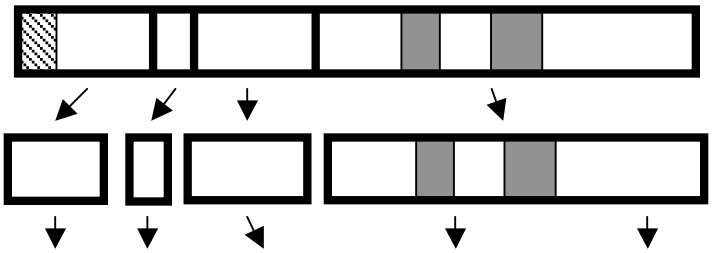


Figure 2.1 Self-interaction of ovulin. Lanes 1, 2, 6, and 7: Loading controls – GST alone (1, 6) or GST-ovulin (2, 7) bound to glutathione beads, probed with anti-GST. Lanes 3 and 4: Pulldowns from male accessory gland extract using GST alone (3) or GST-ovulin (4), probed with anti-ovulin. Lane 5: Accessory gland extract probed with anti-ovulin. The two ovulin bands are different glycosylation forms of the protein, with the upper band at 41 kD and the lower banding consisting of 36 and 37 kD forms (Monsma, Harada, and Wolfner 1990). Lanes 8 and 9: Pulldowns from extracts of female reproductive tracts 30-90 minutes after mating using GST alone (8) or GST-ovulin (9), probed with anti-ovulin. Lane 10: Mated female reproductive tract extract probed with anti-ovulin. The 25 kD and 30 kD bands are consistent in size and timing of appearance with CP3C and an intermediate processing product containing both CP3C and CP3N, respectively.

(a)



(b)



	CP1	CP2	CP3N	CP3C	C45
CP1	-	-	-	-	NA
CP2	-	-	-	-	NA
CP3N	-	-	-	-	NA
CP3C	-	-	-	+	+

Figure 2.2 Interactions among cleavage products of ovulin in yeast two-hybrid assays. Panel (a) AD-CP3C/BD-CP3C diploids grow on His-Trp-Leu- media after 5 days of growth at 30°C, indicating interaction (1), but AD-CP3N/BD-CP3C diploids do not, indicating no interaction (2) (2 streaks of each cell type; other combinations not shown). Panel (b): Summary of two-hybrid interactions. Full length ovulin is depicted at the top, with cleavage products below (Monsma, Harada, and Wolfner 1990). Diagonal lines: Predicted signal sequence. Grey: Regions of sequence similarity to Califin C. Black bars: predicted proteolytic cleavage sites. The + symbol indicates histidine and adenine prototrophy and β-galactosidase activity and the – symbol indicates failure to identify interaction in two-hybrid assays. NA indicates that the assay was not performed.

other putative cleavage products of ovulin (Figure 2.2b), suggesting that the self-interaction is specific**.** Moreover, the failure of the other putative cleavage products to interact with each other, or with CP3C, suggests that CP3C is necessary for self-interaction.

To further delineate the region(s) of ovulin involved in its self-interaction, we tested the ability of the smallest fragment (ovulin C45) that interacted with full length ovulin in the cDNA library screen, to interact with CP3C and with itself. A diploid yeast strain containing CP3C in the DNA binding domain vector and ovulin C45 in the activation domain vector activated the *HIS3* and *lacZ* yeast-two hybrid reporters (Figure 2.2b). Similarly, reporter genes were activated in a diploid strain containing ovulin C45 in both the activation and binding domain vectors (data not shown). Thus, ovulin C45 is sufficient for interaction with CP3C, and interacts with itself, in the yeast two-hybrid system.

To determine whether the C-terminus of ovulin also mediates the interactions we observed in GST-pulldown assays, we performed pulldowns on extracts of reproductive tracts from mated females. GST-ovulin produced *in vitro* successfully pulled down two cleavage products of ovulin from reproductive tract extracts of mated females (Figure 2.1, lanes 6-10). The apparent molecular weights of the pulldown products (25 and 30 kDa; Park and Wolfner 1995), as well as the timing of their appearance (Park and Wolfner 1995), are consistent with those of CP3C, and of an intermediate processing product containing CP3N and CP3C. Thus, our yeast two-hybrid analysis and GST-pulldown assays show that the C-terminus of ovulin is sufficient for self-interaction.

The C-terminus of ovulin contains three potential leucine/isoleucine zippers

We used bioinformatic methods to identify potential structural elements of ovulin responsible for its self-interaction. Secondary structure prediction using PsiPred (McGuffin, Bryson, and Jones 2000) identified three α-helices in CP3C, at amino acids positions 121-136, 143-185, and 212-248. Further manual inspection of these three helices suggested that each contains a potential leucine/isoleucine zipper, at residues 126-136, 171-189, and 229-246 (the two most C-terminal putative zippers are shown in Figure 2.3, a and b). Such zippers are involved in protein-protein interactions, including self-interactions, in a number of proteins, e.g., the HIV protein Vpr and the membrane bound protein phospholamban (Simmerman et al. 1996; Bourbigot et al. 2005). A leucine/isoleucine zipper is thought to consist of an amphipathic helix with a periodicity of 3.5 residues, instead of the usual 3.6, such that residues every two full turns of the helix fall into a straight line. Residues every single full turn (positions 'a' and 'd' on a helical wheel diagram) tend to be occupied by isoleucine or leucine. Further analysis of ovulin using the program COILS identified three candidate coiled-coil domains, corresponding to the putative zippers described above (data not shown). Coiled-coil domains consist of two or more interacting amphipathic α-helices intertwined about each other, and constitute protein-protein interaction interfaces in many polypeptides. The C-terminal putative zipper domain lies within ovulin C45, the smallest fragment identified that is sufficient for self-interaction. We therefore suggest that self-interaction of ovulin is mediated in part by coiled-coil interactions in the C-terminal α-helix of ovulin.

Ovulin participates in SDS-stable complexes

Electrophoresis under non-reducing conditions can yield insights into the oligomerization state of proteins. For example, previous studies have reported that

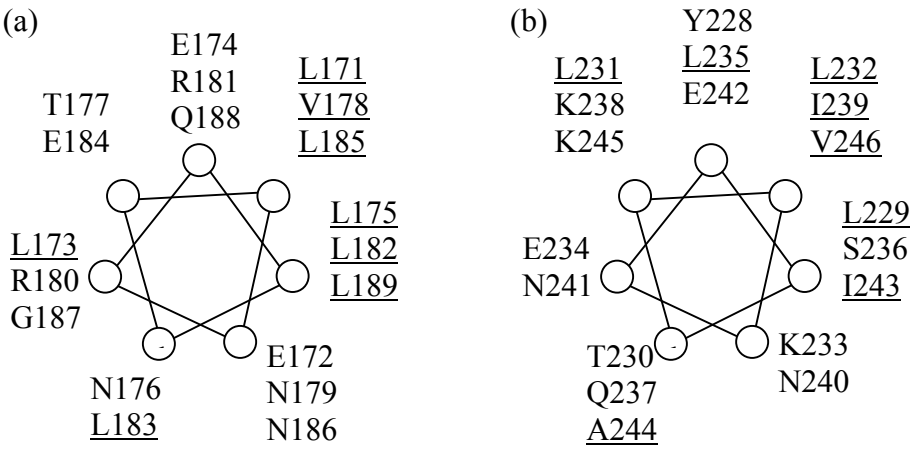


Figure 2.3 Helical wheel diagrams of putative leucine/isoleucine zippers in the C-terminus of ovulin, showing preferential use of leucine and isoleucine residues at positions a and d. Each diagram represents a single α-helix extending into the plane of the paper, with consecutive amino acids arranged alphabetically (a-g) in a clockwise manner. (a) Helical wheel diagram for amino acids 171-189. (b) Helical wheel diagram for amino acids 228-246. Hydrophobic residues are underlined.

pentamers of the leucine/isoleucine zipper protein phospholamban are stable under non-reducing SDS conditions (Wegener and Jones 1984; Simmerman et al. 1996), and some disulfide bonded structures are also resistant to SDS in the absence of reducing agents (e.g., Grigorian et al. 2005; Luo et al. 2005). Thus, we conducted SDS-PAGE of male accessory gland extracts, mated female reproductive tract extracts, and of his_6-tagged ovulin produced in *E. coli*, in the presence or absence of the reducing agent β–mercaptoethanol (β-ME) (Figure 2.4).

In the absence of β-ME, the majority of ovulin present in extracts of male accessory glands and in extracts of mated female reproductive tracts migrates at an apparent molecular weight higher than that predicted and observed for ovulin monomers. The predicted molecular weight of full-length monomeric ovulin is ~30 kD; under reducing conditions, ovulin from the male accessory gland runs as bands of 36-37 kD and 41 kD due to glycosylation (Monsma, Harada, and Wolfner 1990). Under non-reducing conditions, however, ovulin present in the male accessory gland has an apparent molecular weight of 82 kD. In the reproductive tracts of mated females (30-90 minutes post-mating), the major cleavage products CP3C runs at 25 kD under reducing conditions (Park and Wolfner 1995). Under non-reducing conditions, a larger product (55 kD) is again observed from the same sample. Notably, the apparent molecular weights of the bands under non-reducing conditions are consistent with those predicted for dimers of full length ovulin (72-82 kD) or of dimeric CP3C (50 kD). Ovulin thus appears to form homo-oligomers in the male's reproductive tract prior to being transferred to a female. His_6-ovulin produced in *E. coli* also forms higher molecular weight products in the absence of β-ME, although less than half of the recombinant ovulin is found in this larger complex (data not shown). Again, the apparent molecular weight of the higher molecular weight complex is consistent with that predicted for a dimer (60 kD, vs. 30 kD for the non-

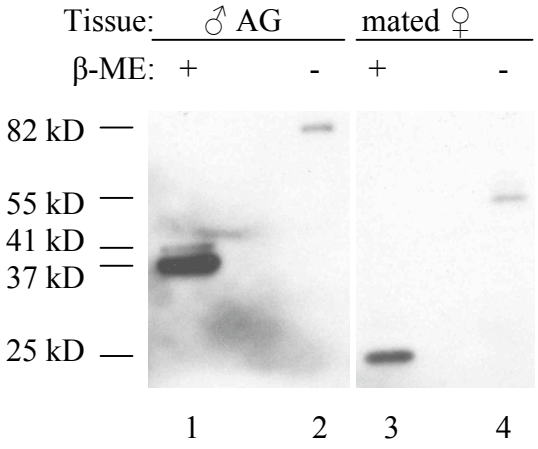
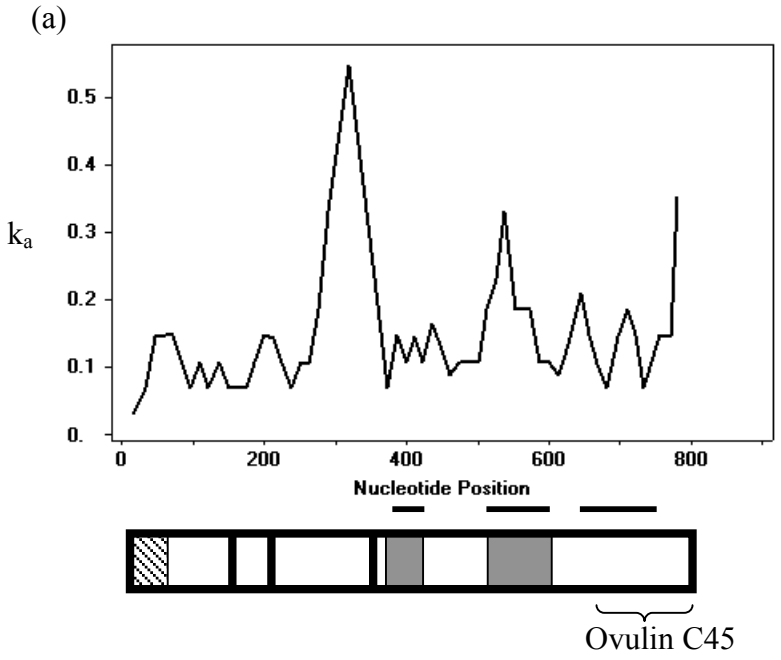


Figure 2.4 Ovulin participates in SDS-stable complexes in extracts from male accessory glands and mated female reproductive tracts. Lanes 1 and 2: Accessory gland extracts in the presence (1) or absence (2) of β-mercaptoethanol (β-ME), probed with anti-ovulin. Lane 1 shows the expected monomeric glycoforms at 36, 37, and 41 kD (Monsma, Harada, and Wolfner 1990), while lane 2 shows a larger, 82 kD product. Lanes 3 and 4: Extracts from the reproductive tracts of females 45-90 minutes post-mating, in the presence (3) or absence (4) of β-ME, probed with anti-ovulin. Lane 3 shows a 25 kD product consistent with CP3C, while lane 4 contains a product of approximately twice the monomeric molecular weight (55 kD).

glycosylated bacterial fusion protein). These results suggest that the formation of ovulin-containing complexes is not dependent on the presence of other seminal fluid proteins.

Conservation of ovulin's putative zipper domains

Ovulin's putative zipper domains are conserved between several species of *Drosophila*, relative to other regions of ovulin (Figure 2.5a). A sliding window plot of the number of nonsynonymous nucleotide substitutions per nonsynonymous site (K_a), a measure of amino acid divergence, shows low levels of sequence divergence in the most C-terminal putative zipper domain (contained within ovulin C45), as well as in the small N-terminal putative zipper at 126-136 (Figure 2.5a). Moreover, leucine and isoleucine residues predicted to be critical to zipper formation are conserved in all three putative zippers between *D. melanogaster* and *D. simulans*, despite 15% amino acid divergence over the entire ovulin protein (the two C-terminal most zippers are shown in Figure 2.5b and c). For the most C-terminal putative zipper domain, the 'a' and 'd' positions are occupied by leucine or isoleucine even in the distantly related species *D. pseudoobscura* (Figure 2.5c), despite only 18.5% overall sequence similarity (Wagstaff and Begun 2005). Moreover, amino acids at the 'g' position of this helix are absolutely conserved at Y228, L235, and E242, perhaps reflecting important inter-molecular interactions (e.g., Harbury, Kim, and Alber 1994). We also noted that C199 is conserved between *D. melanogaster*, *D. simulans*, and *D. pseudoobscura* (data not shown). While this cysteine residue does not lie within ovulin C45, it does fall within CP3C, and may participate in an inter-subunit disulfide bridge. Other conserved residues may also play important structural or functional roles. These comparisons show that putatively critical leucine and isoleucine residues



(b) 171

mel LELELNTVNRRLLELNGQLQNTRKS
sim LELELEALKRKLSEMN--AQNARKP

(c) 229

mel RNKYLTLLKELSQKINNEIAKV---AT
sim RNKYLTLLTQLAQKINYEIANVNNPAT
pse TFRYIRLLRYLRANIDRELTAV----T

Figure 2.5 Conservation of putative zipper regions. Panel (a): Sliding window of nonsynonymous nucleotide divergence between *D. melanogaster* and *D. simulans.* Dark bars below the plot show the locations of the predicted leucine/isoleucine zippers. The schematic diagram of ovulin follows the same color scheme as Figure 2. Panel (b): Protein sequence alignment of second putative zipper (171-189) between *D. melanogaster* (mel) and *D. simulans* (sim). (c) Sequence alignment of third putative zipper (229-246) between *D. melanogaster*, *D. simulans*, and *D. pseudoobscura* (pse). Black arrows denote residues at helix positions a and d, and numbers indicate amino acid position along the *D. melanogaster* protein sequence.

have been conserved in this rapidly diverging protein, consistent with their proposed functional importance.

Discussion

Studies in a variety of animal and plant species have identified numerous rapidly evolving reproductive proteins. A number of explanations have been invoked to account for these observations, including sexual selection (Swanson and Vacquier 2002), defense against pathogens (Mueller et al. 2005), and avoidance of inbreeding and/or self-fertilization (Nasrallah 2002). It is unlikely that any single explanation will account for all cases of rapid evolution in reproductive proteins. Evaluation of a hypothesis with respect to a particular protein or set of proteins requires detailed knowledge of the structural and functional consequences of intra- or inter-specific amino acid variation (e.g., Nasrallah 2002).

Here, we examined the functional architecture of the rapidly evolving *D. melanogaster* egg-laying hormone ovulin. We found that ovulin interacts with itself, and that its C-terminal most cleavage product, CP3C, is sufficient for this interaction. Furthermore, yeast two-hybrid analyses suggest that the last 45 amino acids of ovulin (ovulin C45) are sufficient for self-interaction. CP3C contains three potential leucine/isoleucine zipper regions which are predicted to form coiled-coils; one of these regions falls within ovulin C45. Zipper domains are involved in dimerization or oligomerization in a number of systems (Simmerman et al. 1996; Bourbigot et al. 2005; Takemoto and Hibi 2005). Notably, the proposed self-interaction domain in ovulin C45 shows a high degree of conservation at potentially crucial residues, despite ovulin's overall rapid protein divergence.

While the functional significance of ovulin's self-interaction is not yet known, a number of other prohormones and hormones are known to dimerize. For example,

the *Bombyx mori* neuropeptide prothoracicotropic hormone (PTTH) (Kataoka et al. 1991) and the locust arginine-vasopression-like diuretic hormone F2 (Proux et al. 1987) both form homodimers. In the latter case, the bioactivity of dimeric F2 is greatly enhanced over that of its monomeric counterpart, F1. In *D. melanogaster*, dimerization is also necessary for the function of the heterodimeric cuticle hardening hormone bursicon (Luo et al. 2005), and human transforming growth factor-β3 (TGF-β3) binds its receptor as a dimer (Hart et al. 2002). By analogy, self-interaction may also be essential for ovulin's action. Tests of this hypothesis, while desirable, present experimental and interpretive challenges at this time.

The conservation of critical leucine and isoleucine residues in ovulin's proposed coiled-coil domains suggests that this protein's self-interaction may in fact be important for functionality. In other rapidly evolving proteins, conserved structural domains or residues are thought to play important structural roles. In the vertebrate MHC class I molecule, for example, rapidly evolving residues localize to the antigen recognition site (Hughes and Nei 1988). Other portions of the molecule, including the macroglobulin binding α_3 domain and the single transmembrane helix, show much lower rates of sequence evolution. Similarly, a few key residues appear to maintain tertiary structure in alleles of the crucifer SCR protein (Chookajorn et al. 2004).

In light of these examples, we suggest that ovulin's self-interaction, and specifically the protein domain(s) involved in its self-interaction, are important for ovulin's function. Constraint on a few key structural elements in the self-interaction domain may reduce constraint on primary sequence and local secondary/tertiary structure elsewhere in the protein, thereby providing conditions that would allow rapid adaptive evolution.

REFERENCES

Adams, M. D., S. E. Celniker, R. A. Holt, C. A. Evans, J . D. Gocayne, P. G. Amanatides, S. E. Scherer, P. W. Li, R. A. Hoskins, R. F. Galle *et al.* 2000. The genome sequence of *Drosophila melanogaster*. Science **287**:2185-2195.

Aguadé, M. 1998. Different forces drive the evolution of the Acp26Aa and Acp26Ab accessory gland genes in the *Drosophila melanogaster* species complex. Genetics **150**:1079-1089.

Aguadé, M., N. Miyashita, and C. H. Langley. 1992. Polymorphism and divergence in the Mst26A male accessory gland gene region in Drosophila. Genetics **132**:755-770.

Andolfatto, P. 2005. Adaptive evolution of non-coding DNA in Drosophila. Nature **437**:1149-1152.

Bendtsen, J. D., H. Nielsen, G. von Heijne, and S. Brunak. 2004. Improved prediction of signal peptides: SignalP 3.0. J Mol Biol **340**:783-795.

Bloch Qazi, M. C., Y. Heifetz, and M. F. Wolfner. 2003. The developments between gametogenesis and fertilization: ovulation and female sperm storage in *Drosophila melanogaster*. Dev Biol **256**:195-211.

Bloom, J. D., S. T. Labthavikul, C. R. Otey, and F. H. Arnold. 2006. Protein stability promotes evolvability. Proc Natl Acad Sci U S A **103**:5869-5874.

Bourbigot, S., H. Beltz, J. Denis, N. Morellet, B. P. Roques, Y. Mely, and S. Bouaziz. 2005. The C-terminal domain of the HIV-1 regulatory protein Vpr adopts an antiparallel dimeric structure in solution via its leucine-zipper-like domain. Biochem J **387**:333-341.

Chapman, T., and S. J. Davies. 2004. Functions and analysis of the seminal fluid proteins of male *Drosophila melanogaster* fruit flies. Peptides **25**:1477-1490.

Chookajorn, T., A. Kachroo, D. R. Ripoll, A. G. Clark, and J. B. Nasrallah. 2004. Specificity determinants and diversification of the Brassica self-incompatibility pollen ligand. Proc Natl Acad Sci U S A **101**:911-917.

Clark, A. G., M. Aguadé, T. Prout, L. G. Harshman, and C. H. Langley. 1995. Variation in sperm displacement and its association with accessory gland protein loci in *Drosophila melanogaster*. Genetics **139**:189-201.

DeLange, R. J., D. M. Fambrough, E. L. Smith, and J. Bonner. 1969. Calf and pea histone IV. 3. Complete amino acid sequence of pea seedling histone IV; comparison with the homologous calf thymus histone. J Biol Chem **244**:5669-5679.

Derynck, R., J. A. Jarrett, E. Y. Chen, D. H. Eaton, J. R. Bell, R. K. Assoian, A. B. Roberts, M. B. Sporn, and D. V. Goeddel. 1985. Human transforming growth factor-beta complementary DNA sequence and expression in normal and transformed cells. Nature **316**:701-705.

Fiumera, A. C., B. L. Dumont, and A. G. Clark. 2005. Sperm competitive ability in *Drosophila melanogaster* associated with variation in male reproductive proteins. Genetics **169**:243-257.

Gillott, C. 2003. Male accessory gland secretions: modulators of female reproductive physiology and behavior. Annu Rev Entomol **48**:163-184.

Grigorian, A. L., J. J. Bustamante, P. Hernandez, A. O. Martinez, and L. S. Haro. 2005. Extraordinarily stable disulfide-linked homodimer of human growth hormone. Protein Sci **14**:902-913.

Harbury, P. B., P. S. Kim, and T. Alber. 1994. Crystal structure of an isoleucine-zipper trimer. Nature **371**:80-83.

Hart, P. J., S. Deep, A. B. Taylor, Z. Shu, C. S. Hinck, and A. P. Hinck. 2002. Crystal structure of the human TbetaR2 ectodomain--TGF-beta3 complex. Nat Struct Biol **9**:203-208.

Heifetz, Y., O. Lung, E. A. Frongillo, Jr., and M. F. Wolfner. 2000. The Drosophila seminal fluid protein Acp26Aa stimulates release of oocytes by the ovary. Curr Biol **10**:99-102.

Heifetz, Y., L. N. Vandenberg, H. I. Cohn, and M. F. Wolfner. 2005. Two cleavage products of the Drosophila accessory gland protein ovulin can independently induce ovulation. Proc Natl Acad Sci U S A **102**:743-748.

Herndon, L. A., and M. F. Wolfner. 1995. A Drosophila seminal fluid protein, Acp26Aa, stimulates egg laying in females for 1 day after mating. Proc Natl Acad Sci U S A **92**:10114-10118.

Hirokawa, T., S. Boon-Chieng, and S. Mitaku. 1998. SOSUI: classification and secondary structure prediction system for membrane proteins. Bioinformatics **14**:378-379.

Hughes, A. L., and M. Nei. 1988. Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. Nature **335**:167-170.

Kataoka, H., H. Nagasawa, A. Isogai, H. Ishizaki, and A. Suzuki. 1991. Prothoracicotropic hormone of the silkworm, *Bombyx mori*: amino acid sequence and dimeric structure. Agric Biol Chem **55**:73-86.

Kopp, A., and J. R. True. 2002. Evolution of male sexual characters in the oriental Drosophila melanogaster species group. Evol Dev **4**:278-291.

Kubli, E. 2003. Sex-peptides: seminal peptides of the Drosophila male. Cell Mol Life Sci **60**:1689-1704.

Luo, C. W., E. M. Dewey, S. Sudo, J. Ewer, S. Y. Hsu, H. W. Honegger, and A. J. Hsueh. 2005. Bursicon, the insect cuticle-hardening hormone, is a heterodimeric cystine knot protein that activates G protein-coupled receptor LGR2. Proc Natl Acad Sci U S A **102**:2820-2825.

Lupas, A., M. Van Dyke, and J. Stock. 1991. Predicting coiled coils from protein sequences. Science **252**:1162-1164.

McGuffin, L. J., K. Bryson, and D. T. Jones. 2000. The PSIPRED protein structure prediction server. Bioinformatics **16**:404-405.

Monsma, S. A., H. A. Harada, and M. F. Wolfner. 1990. Synthesis of two Drosophila male accessory gland proteins and their fate after transfer to the female during mating. Dev Biol **142**:465-475.

Monsma, S. A., and M. F. Wolfner. 1988. Structure and expression of a Drosophila male accessory gland gene whose product resembles a peptide pheromone precursor. Genes Dev **2**:1063-1073.

Mueller, J. L., K. R. Ram, L. A. McGraw, M. C. Bloch Qazi, E. D. Siggia, A. G. Clark, C. F. Aquadro, and M. F. Wolfner. 2005. Cross-species comparison of Drosophila male accessory gland protein genes. Genetics **171**:131-143.

Nasrallah, J. B. 2002. Recognition and rejection of self in plant reproduction. Science **296**:305-308.

Park, M., and M. F. Wolfner. 1995. Male and female cooperate in the prohormone-like processing of a *Drosophila melanogaster* seminal fluid protein. Dev Biol **171**:694-702.

Pitnick, S., T. A. Markow, and G. S. Spicer. 1995. Delayed male maturity is a cost of producing large sperm in Drosophila. Proc Natl Acad Sci U S A **92**:10614-10618.

Proux, J. P., C. A. Miller, J. P. Li, R. L. Carney, A. Girardie, M. Delaage, and D. A. Schooley. 1987. Identification of an arginine vasopressin-like diuretic hormone from Locusta migratoria. Biochem Biophys Res Commun **149**:180-186.

Rozas, J., J. C. Sanchez-DelBarrio, X. Messeguer, and R. Rozas. 2003. DnaSP, DNA polymorphism analyses by the coalescent and other methods. Bioinformatics **19**:2496-2497.

Seidah, N. G., S. Benjannet, J. Hamelin, A. M. Mamarbachi, A. Basak, J. Marcinkiewicz, M. Mbikay, M. Chretien, and M. Marcinkiewicz. 1999. The subtilisin/kexin family of precursor convertases. Emphasis on PC1, PC2/7B2, POMC and the novel enzyme SKI-1. Ann N Y Acad Sci **885**:57-74.

Simmerman, H. K., Y. M. Kobayashi, J. M. Autry, and L. R. Jones. 1996. A leucine zipper stabilizes the pentameric membrane domain of phospholamban and forms a coiled-coil pore structure. J Biol Chem **271**:5941-5946.

Swaffield, J. C., and S. A. Johnston. 1996. Affinity purification of proteins binding to GST fusion proteins *in* F. M. Ausubel, R. Brent, R. E. Kingston, D. D. Moore, J. G. Seidman, J. A. Smith, and K. Struhl, eds. Current Protocols in Molecular Biology. John Wiley and Sons, Inc., New York.

Swanson, W. J., and V. D. Vacquier. 2002. The rapid evolution of reproductive proteins. Nat Rev Genet **3**:137-144.

Takemoto, Y., and T. Hibi. 2005. Self-interaction of ORF II protein through the leucine zipper is essential for Soybean chlorotic mottle virus infectivity. Virology **332**:199-205.

Tamura, K., S. Subramanian, and S. Kumar. 2004. Temporal patterns of fruit fly (Drosophila) evolution revealed by mutation clocks. Mol Biol Evol **21**:36-44.

Tsaur, S. C., C. T. Ting, and C. I. Wu. 1998. Positive selection driving the evolution of a gene of male reproduction, Acp26Aa, of Drosophila: II. Divergence versus polymorphism. Mol Biol Evol **15**:1040-1046.

Tsaur, S. C., and C. I. Wu. 1997. Positive selection and the molecular evolution of a gene of male reproduction, Acp26Aa of Drosophila. Mol Biol Evol **14**:544-549.

Tusnady, G. E., and I. Simon. 2001. The HMMTOP transmembrane topology prediction server. Bioinformatics **17**:849-850.

Wagstaff, B. J., and D. J. Begun. 2005. Comparative genomics of accessory gland protein genes in *Drosophila melanogaster* and *D. pseudoobscura*. Mol Biol Evol **22**:818-832.

Wegener, A. D., and L. R. Jones. 1984. Phosphorylation-induced mobility shift in phospholamban in sodium dodecyl sulfate-polyacrylamide gels. Evidence for a protein structure consisting of multiple identical phosphorylatable subunits. J Biol Chem **259**:1834-1841.

Wolfner, M. F. 2002. The gifts that keep on giving: physiological functions and evolutionary dynamics of male seminal proteins in Drosophila. Heredity **88**:85-93.

Wolfner, M. F., Y. Heifetz, and S. W. Applebaum. 2005. Gonadal glands and their gene products *in* L. I. Gilbert, K. Iatrou, and S. S. Gill, eds. Comprehensive molecular insect science: Reproduction and Development. Elsevier Ltd., Oxford.

CHAPTER 3

IMMORTAL COILS: CONSERVED DIMERIZATION MOTIFS OF THE DROSOPHILA EGG-LAYING PROHORMONE OVULIN

Introduction

Recently, much attention has been paid to the evolvability of biological systems. While definitions of evolvability vary widely between authors (Sniegowski and Murphy 2006; Pigliucci 2008), an intuitive and interesting conception of evolvability is a system's ability to withstand new variation that may be favored by selection. Authors have sought to describe and understand the evolvability of a wide range of biological entities, including phylogenetic lineages (Kirschner and Gerhart 1998), regulatory networks (Kirschner and Gerhart 1998), and proteins (Wilke et al. 2005; Bloom et al. 2006).

Proteins can differ substantially in their abilities to withstand sequence variation. At one extreme, the conservation of proteins such as core histones (DeLange et al. 1969) and β-tubulin (Kirschner and Gerhart 1998) indicates that most mutations are highly deleterious, such that essentially no variation can be tolerated. Other proteins, by contrast, are tolerant of substantial sequence diversity, with primary sequence relatively unimportant for protein function. In some cases, variation may be favored, with positive selection maintaining polymorphisms within species and/or fixing substitutions between species. Extremely high polymorphism in parts of the vertebrate class I MHC molecules (Hughes and Nei 1988), for example, is thought be driven by a molecular 'arms race' with pathogens. While tolerance to variation does not imply the action of positive selection (variation could be neutral), selection does require a certain amount of tolerance – there must be variation upon which selection can act.

A number of factors have been proposed to contribute to a protein's evolvability. External constraints may be imposed on proteins by, for example, their interactions with other proteins, their expression in multiple tissues, translational considerations, or by their involvement in multiple processes, leading to reduced overall rates of evolution and/or a decreased propensity to undergo positive selection (e.g., Drummond, Raval, and Wilke 2006; Larracuente et al. 2008). Features of proteins themselves may also promote evolvability (Bloom et al. 2005b; Wilke et al. 2005; Bloom et al. 2006). A protein's stability has been proposed to be an important contributor, since a more stable protein will be better able to withstand the potentially destabilizing effects of new mutations (Bloom et al. 2006). Nonetheless, the role of a protein's structural features in determining its evolvability remains relatively unexplored; do certain features reduce intrinsic constraints on a protein's function?

Rapidly evolving proteins provide good empirical examples with which to study the determinants of evolvability, since such molecules can clearly withstand substantial sequence variation. Here, we focus on the rapidly evolving Drosophila seminal protein ovulin. Ovulin is a prohormone synthesized in the male's accessory glands, secretory structures in the reproductive tract that produce a substantial fraction of the seminal fluid. During mating, ovulin is transferred to the female along with sperm and >100 other accessory gland proteins (Acps; Ravi Ram and Wolfner 2007). Ovulin increases a female's rate of ovulation by ~10-20% for about a day after mating (Monsma and Wolfner 1988; Herndon and Wolfner 1995; Heifetz et al. 2005). Ovulin is proteolytically cleaved into four smaller products following mating (Park and Wolfner 1995; Heifetz et al. 2005), with at least one other Acp, the protease CG11864, necessary for cleavage (Ravi Ram, Sirot, and Wolfner 2006). Each of ovulin's two C-terminal cleavage fragments, as well as the full-length protein, are individually

sufficient to induce ovulation upon ectopic expression (Heifetz et al. 2005), suggesting some degeneracy of function.

Ovulin's tolerance to mutations is clearly demonstrated by its tremendous diversity within and between species. For example, ovulin's 15% amino acid divergence between the closely related species *D. melanogaster* and *D. simulans* far exceeds the genome wide average of ~1-2%. Moreover, at least in some species, polymorphism at the *ovulin* locus is elevated; in *D. mauritiana*, for example, polymorphism is substantially higher than at other loci in this species (Tsaur, Ting, and Wu 2001). A number of studies have provided strong evidence that ovulin's high levels of polymorphism and divergence are due, at least in part, to positive selection (Aguadé, Miyashita, and Langley 1992; Tsaur and Wu 1997; Aguadé 1998; Tsaur, Ting, and Wu 1998). This positive selection is likely the result of some form of sexual selection, e.g., sexual conflict (Aguadé, Miyashita, and Langley 1992; Swanson and Vacquier 2002; Wong, Albright, and Wolfner 2006).

Previously, we showed that ovulin self-interacts, likely occurring as a dimer (or other multimer). We hypothesized that one or more putative coiled-coils in ovulin's C-terminal cleavage fragment mediate this self-interaction (Chapter 2; Wong, Albright, and Wolfner 2006). Interestingly, ovulin's potential coiled-coils are highly conserved between species, as they are identifiable even in *D. pseudoobscura*, a distant relative of *D. melanogaster* whose ovulin is ~80% divergent overall from its *D. melanogaster* ortholog at the amino acid level (Wagstaff and Begun 2005). We proposed that constraint on these few structural motifs might reduce constraint on other portions of the protein, contributing to its apparently high evolvability.

In this study, we provide evidence that at least one of ovulin's putative coiled-coils, as well as a conserved tyrosine motif, are indeed necessary for self-interaction, and that ovulin's dimeric structure is conserved between species despite considerable

sequence divergence. Thus, consistent with our earlier hypotheses, aspects of ovulin's tertiary structure are robust to high levels of sequence variation. We also demonstrate that putative coiled-coil proteins in the Drosophila genome tend to evolve more rapidly than proteins predicted to lack coiled-coils, suggesting that such structures may contribute widely to protein evolvability.

Materials and Methods

Fly rearing and analysis of accessory gland extracts

Strains of *D. melanogaster* (Canton-S) and *D. simulans* (Sim6) were maintained on yeast-glucose media at room temperature on 12 hour light:12 hour dark cycles. For analysis of SDS-stable ovulin complexes (Wong, Albright, and Wolfner 2006), the accessory glands of 4-7 day old males were dissected into 20 μl 40% sucrose + protease inhibitor cocktail (Roche; Indianapolis, IN) and homogenized with a pestle. 10 μl of SDS-PAGE loading buffer with 0.1% β-mercaptoethanol (β-ME) was added to half of the resulting mixture, while 10 μl SDS-PAGE loading buffer without β-ME was added to other half. Samples were separated by 10% SDS-PAGE, and Western blotting was performed using α-ovulin antibodies at a 1:2000 concentration, followed by α-rabbit IgG secondary antibodies at 1:2000.

Cloning, protein synthesis, and cross-linking

3' fragments of the *ovulin* gene were amplified by polymerase chain reaction (PCR) from *D. melanogaster* (aa 219-264 of 264), *D. simulans* (aa 156-255 of 255), *D. yakuba* (aa 179-234 of 234), and *D. pseudoobscura* (aa 209-247 of 247), and cloned in the Gateway compatible entry vector pENTR-dTopo (Invitrogen; Carlsbad, CA) according the manufacturer's instructions. Site-directed mutagenesis of the *D. melanogaster* entry clone was performed using the QuikChange kit (Stratagene; Cedar

Creek, TX). Expression clones in vector pEXP1-DEST (Invitrogen) were generated by LR reaction (Invitrogen).

In vitro protein synthesis was performed using the Expressway Mini Cellfree Expression System (Invitrogen) according the manufacturer's protocol. Cross-linking was performed by adding 10 μl of 10mM dimethyl suberimidate•2 HCl (DMS) in phosphate buffered saline pH 8.0 (PBS) to 10 μl of crude *in vitro* protein synthesis mixture and incubating for 1 hour at room temperature. DMS consists of two reactive imidoester groups separated by a 11.0 angstrom spacer arm; the imidoester groups react with amine groups, found primarily on lysine side chains and the N-termini of proteins, to form covalently cross-linked structures. We used DMS (rather than non-reducing conditions) to examine self-interaction of C-terminal ovulin fragments since preliminary data (not shown) suggested that dimers were not SDS-stable, perhaps due to the absence of Cys199 in these peptides. Controls were performed using 10 μl of PBS without DMS. 20 μl of SDS-PAGE sample buffer was added to stop the reaction. Proteins were separated by SDS-PAGE, and detected by Western blotting using α-His antibodies (Sigma) at a 1:2500 concentration.

Analysis of rates of substitution

Estimates of *dN* and ω and inferences of positive selection were performed using PAML (Yang 2007) by (Larracuente et al. 2008) for 8510 genes (not including ovulin) with clear one-to-one orthologs in 6 species of the genus Drosophila: *D. melanogaster*, *D. simulans*, *D. sechellia*, D. *yakuba*, *D. erecta*, and *D. ananassae.* Briefly, gene-averaged *dN* and ω were estimated under model M0, which assumes a single value of *dN* and *dS* for each gene across the entire phylogeny. For inferences of positive selection, model M7 was used as the null hypothesis, allowing beta-distributed variation in ω but disallowing codons with $\omega > 1$. The alternative

hypothesis M8 also allows beta-distributed variation in ω, and adds a class of codons with ω > 1. For each gene, a likelihood ratio test can be used to determine if the data fit model M8 better than they fit model M7; a rejection of M7 constitutes evidence in favor of recurrent positive selection on a subset of codons. False discovery rate (FDR) corrections were performed as described in Larracuente et al. 2008.

We examined the influence of several factors on *dN*, ω, and the likelihood of positive selection: Presence/absence of putative coiled-coil domains, protein length, tissue specificity, maximum expression level, length and number of introns, and local recombination rate. The presence of one or more coiled-coil domains was predicted using PairCoil (McDonnell et al. 2006), using default parameters. All other factors are described in Larracuente et al. 2008. We used multiple linear regression in order to infer the contribution of each factor to *dN* and ω, both of which are continuous variables. Logistic regression was used to infer the contribution of each factor to positive selection, where the outcome for each gene is binary (selected or not selected). Statistical analyses were performed using R version 2.5.1 (R Core Development Team 2008).

Results

Conserved motifs of ovulin

Previously, we showed that ovulin's 147 aa C-terminal cleavage product, CP3C, is sufficient for self-interaction. Moreover, within CP3C, the C-terminal 45 amino acids of ovulin are capable of self-interaction in yeast two-hybrid assays. These findings prompted us to look for potential interaction motifs in ovulin's C-terminus. We have identified three such motifs, all of which are highly conserved between species (Figure 3.1A): (1) Three potential coiled-coil motifs are present in CP3C, as described in Chapter 2 and Wong et al. (2006). Coiled-coils consist of two or more

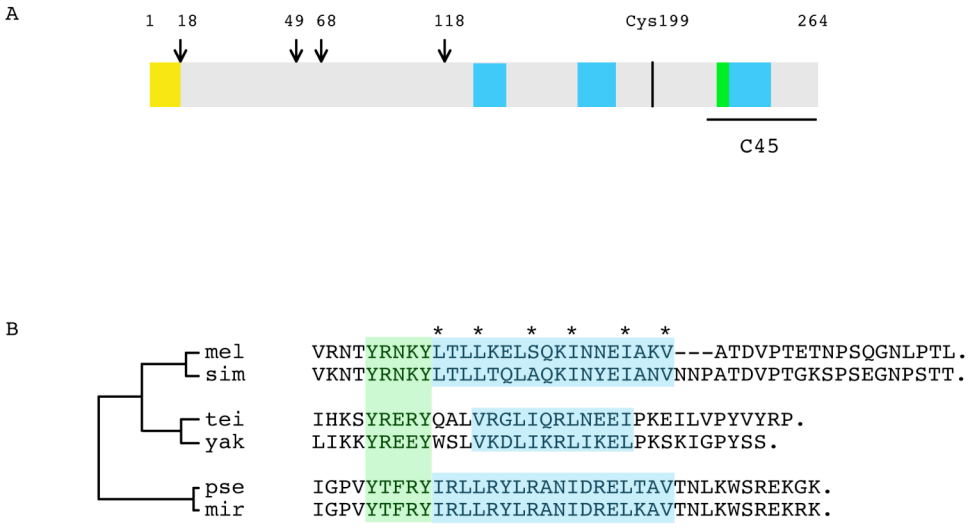


Figure 3.1 Conserved putative interaction motifs of ovulin. Panel A: Schematic of *D. melanogaster* ovulin. Three kinds of motif are hypothesized to contribute to self-interaction: Putative coiled-coils (blue), a YxxxY motif (green), and ^{199}Cys. The putative secretion signal sequence is indicated in yellow, and putative proteolytic cleavage sites are marked with arrows. The C-terminal 45 amino acids of ovulin (C45) were previously shown to be sufficient for self-interaction in yeast two-hybrid assays (Wong et al. 2006; Chapter 2). Panel B: Predicted amino acid sequence of *D. melanogaster* C45 (mel) and orthologous regions from *D. simulans* (sim), *D. teissieri* (tei), *D. yakuba* (yak), *D. pseudoobscura* (pse), and *D. miranda* (mir). This fragment of ovulin contains the YxxxY motif (green) and one putative coiled-coil (blue). Residues predicted to form the hydrophobic face of the coiled-coil are indicated with asterisks. Branch lengths on the phylogenetic tree are for illustrative purposes only (not to scale).

interacting α-helices, each of which has a strongly hydrophobic face. Participating α-helices wrap around each other by virtue of hydrophobic interactions between these surfaces. The hallmark of a coiled-coil, then, is an α-helix with hydrophobic amino acids occurring every 3-4 residues. (2) ^{199}Cys may participate in an inter-subunit disulfide bond. Consistent with a role for a disulfide bond in dimerization, putative ovulin dimers are SDS stable in the absence of reducing agent (Chapter 2; Wong et al. 2006). (3) A YxxxY motif may also contribute to self-interaction. A similar YxxxY motif is known to mediate interactions between thymine DNA glycosylase (TDG) and SRC1 (Lucey et al. 2005); this finding, combined with the conservation of this motif (see below), led us to consider ovulin's YxxxY motif as a potential self-interaction sequence.

The two candidate self-interaction motifs present in C45, YxxxY and one coiled-coil, show remarkable conservation in this otherwise rapidly evolving protein (Figure 3.1B). Both tyrosines in the YxxxY motif are absolutely conserved between *D. melanogaster* and *D. pseudoobscura*, and there also appears to be a tendency for intervening residues to carry a charge. In particular, the third intervening residue is charged in all species for which sequences are available, spanning ~25 million years of evolution and ~80% amino acid divergence across the ovulin protein. The YxxxY motifs of TDG and SRC1, by contrast, do not appear to have a strong requirement for charged residues (Lucey et al. 2005).

Conserved coiled-coil and YxxxY motifs are necessary for the self-interaction of C45 in vitro

We used an *in vitro* synthesis and cross-linking approach to test the roles of the conserved YxxxY and coiled-coil motifs in the self-interaction of C45. Wild-type C45 tagged at its N-terminus with 6xHis and Xpress (Invitrogen) epitopes, or mutants

bearing alterations of residues with predicted roles in self-interaction (Figure 3.2A), were produced using a cell-free *E. coli* extract (Invitrogen), and the products were either cross-linked using DMS or incubated in buffer alone. Wild-type C45 incubated in buffer alone runs at about 9 kD on an SDS-PAGE gel, consistent with a monomeric peptide; a product of ~18 kD is present when wild-type C45 is cross-linked with DMS (Figure 3.2B), consistent with the presence of cross-linked dimers.

We made three mutants that are predicted to disrupt the coiled-coil: Zip1, Zip2, and Zip3 each bear two alanine mutations of residues predicted to lie within the hydrophobic core of the coiled-coil (Figure 3.2A). Each of the coiled-coil mutants is predicted to disrupt the coiled-coil's hydrophobic face through two turns of a single α-helix. Higher molecular weight products were not detectable following DMS treatment for any of these three mutants, suggesting that they are incapable of forming dimers (results for Zip1 and Zip2 are shown in Figure 3.2B; comparable results were obtained for Zip3 – data not shown). This result is consistent with a coiled-coil interaction interface between monomers of C45.

Similarly, mutations to the YxxxY motif greatly reduce or eliminate C45's ability to self-interact. No dimer is detected following DMS treatment when either or both tyrosines in the YxxxY motif are converted to alanine (Figure 3.2B; results shown for the double YYAA mutant only). Interestingly, a ^{224}Y->F mutation also abrogates self-interaction (data not shown), suggesting that other bulky aromatic residues cannot substitute for tyrosine in this motif; similar results were obtained by (Lucey et al. 2005) for the TDG/SRC1 YxxxY motif. We attempted to assay self-interaction of a ^{228}Y->F mutant and of a double Y->F mutant, but protein yields were very low. It is unclear at this point whether ovulin's YxxxY motif is structurally similar to those of TDG and SRC1; in the latter case, the YxxxY motif is repeated several times, while ovulin bears only one.

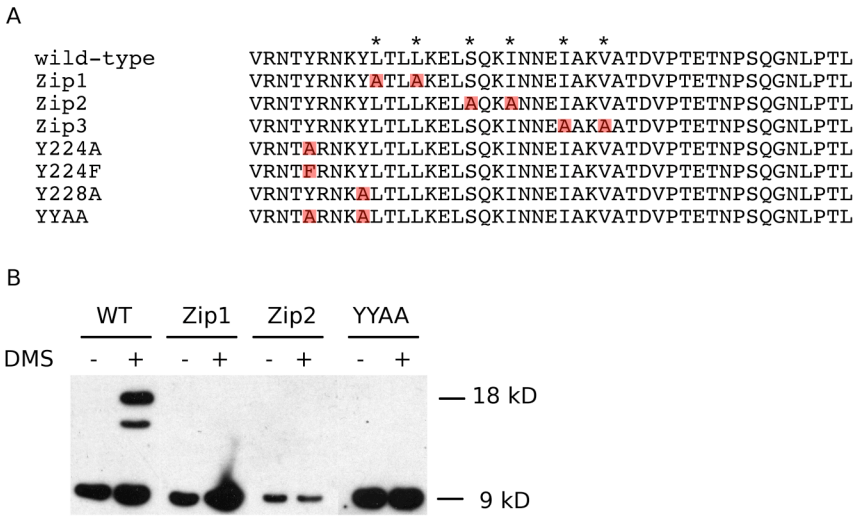


Figure 3.2 Conserved motifs in ovulin C45 are necessary for dimerization. Panel A: Amino acid sequence of *D. melanogaster* C45 and of seven mutants assayed for self-interaction. Residues predicted to form the hydrophobic face of the coiled-coil are indicated with asterisks. Mutated residues are highlighted in red. B: Western blots of wild-type and mutant forms of C45 produced *in vitro*, after exposure to the cross-linker DMS (+) or buffer only (-). Wild-type C45 forms SDS/β-ME stable products in the presence of DMS, the molecular weight of which is consistent with a dimer (a smaller product may be a degradation product). Coiled-coil mutants (Zip1 and Zip2) and a mutant with Y->A mutations in the YxxxY motif (YYAA) fail to form dimers in the presence of DMS.

Self-interaction of putative ovulin orthologs from other species of Drosophila suggests a conserved role for interaction motifs

Given the conservation of all three putative self-interaction motifs in ovulin, we predicted that ovulin orthologs from other species of Drosophila should also self-interact. In order to test this hypothesis, we prepared accessory gland extracts from *D. melanogaster* and *D. simulans*, and performed SDS-PAGE in the presence or absence of the reducing agent β-ME (Figure 3.3A). As observed previously, *D. melanogaster* ovulin migrates as two bands of 37 kD and 41 kD under reducing conditions, representing different glycosylation products (Monsma, Harada, and Wolfner 1990). In the absence of β-ME, however, it runs at ~80 kD, consistent with a disulfide bonded dimer. At least one other known coiled-coil protein exhibits similar behavior (Simmerman et al. 1996). We obtained similar results for the putative ovulin ortholog of *D. simulans* (Figure 3.3A), consistent with self-interaction of ovulin in this species. Results with other species of Drosophila (*D. yakuba*, *D. teissieri, D. takahashii*) were inconclusive; high protein divergence appeared to make recognition using *D. melanogaster* α-ovulin antibodies difficult.

We predicted that C-terminal fragments of ovulin from non-*melanogaster* species of *Drosophila* should be capable of self-interaction, given the concentration of conserved interaction motifs in ovulin's C-terminus. We therefore expressed C-terminal peptides of ovulin from *D. simulans*, *D. yakuba*, and *D. pseudoobscura* in cell free *E. coli* extracts and assayed self-interaction by DMS cross-linking (Figure 3.3B). Note that, for this experiment, peptides were detected using an antibody against a 6xHis tag, such that divergence does not affect detection. As predicted, products whose molecular weight is consistent with a dimer were present following cross-linking for all species assayed. Different cross-linking efficiency was observed for

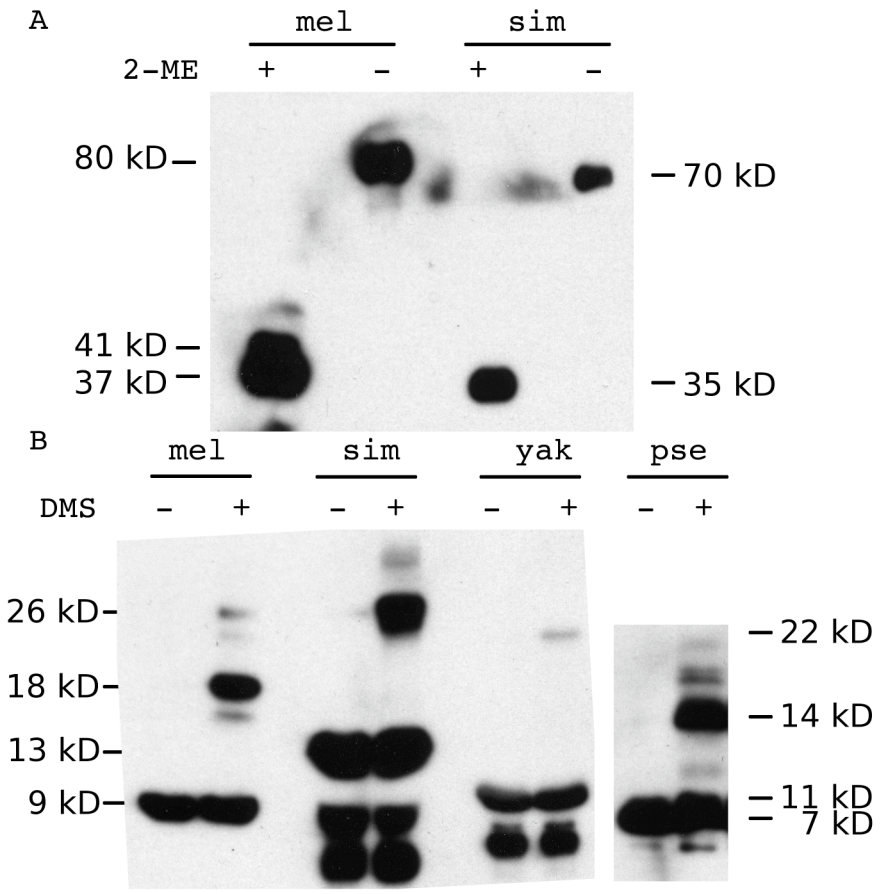


Figure 3.3 Putative ovulin complexes in non-*melanogaster* species. Panel A: Western blots using anti-ovulin antibodies on accessory gland extracts from *D. melanogaster* (mel) and *D. simulans* (sim) in the presence (+) or absence (-) of the reducing agent 2-ME. Panel B: Western blots of *D. melanogaster* C45 and C-terminal ovulin fragments from *D. simulans*, *D. yakuba*, and *D. pseudoobscura* produced *in vitro*, following treatment with buffer alone (-) or with DMS (+). Expected monomeric molecular weights are 9 kD, 13 kD, 11 kD, and 7 kD, respectively. Note the presence of a product consistent with a dimer in all species following treatment with DMS. Extra bands below 13 kD (*D. simulans*) or 11 kD (*D. pseudoobscura*) may be degradation products.

different species; in particular, the putative dimer signal is relatively weak for *D. yakuba*. This may reflect differences in the strengths of inter-subunit associations; alternatively, different lysine content and positioning for different species could contribute to differences in cross-linking efficiency regardless of bond strengths.

Interestingly, yeast-two hybrid analyses suggest that *D. melanogaster* ovulin and its orthologs in *D. simulans*, *D. sechellia*, and *D. mauritiana* are capable of forming heterospecific dimers (Adam Christopher, unpublished data). Thus, it appears that the tertiary structure of ovulin is sufficiently conserved to allow dimerization even in the face of 15% amino acid divergence protein-wide.

These results suggest that similar tertiary structures are adopted by the putative coiled-coil domains of ovulin orthologs from a variety of species, despite an overall high rate of protein divergence. Motifs shown to be essential for dimerization in *D. melanogaster* are conserved in species as distant as *D. pseudoobscura*, and ovulin orthologs in multiple species appear to form dimers. Thus, ovulin's dimeric structure is robust to very high levels of primary sequence divergence.

Evidence for reduced constraint on predicted coiled-coil proteins

If ovulin's conserved dimerization motifs impose relatively few sequence constraints, and thereby contribute to ovulin's evolvability, then we predicted that other proteins bearing similar domains should also show reduced constraint, as indicated by an increased rate of amino acid evolution (*dN*), and an increased *dN*/*dS* ratio (ω). We have tested this hypothesis using the recently reported sequences of the genomes of 12 species of Drosophila (Clark et al. 2007). Larracuente et al. (2008) estimated evolutionary rates and tested for positive selection at 8510 genes with clear one-to-one orthologs in *D. melanogaster* and its five closest sequenced relatives, *D. simulans*, *D. sechellia*, *D. yakuba*, *D. erecta*, and *D. ananassae.* Here, we focus on the

evolutionary dynamics of predicted coiled-coil proteins, since coiled-coils are well characterized protein-protein interaction motifs and can be reliably identified using computational methods (e.g., McDonnell et al. 2006). We note that many coiled-coils participate in non-self interactions (e.g., Vinson, Acharya, and Taparowsky 2006); this is not problematic for our analysis, since coiled-coils still fulfill important structural roles, whether in self- or non-self-interactions.

We compared rates of nucleotide divergence between proteins containing predicted coiled-coils and those without predicted coiled-coils (Table 3.1). After controlling for several other parameters previously demonstrated to contribute to variation in the rate of amino acid substitution (tissue specificity, expression level, protein length, intron number and length, recombination rate - Larracuente et al. 2008), we found that predicted coiled-coil proteins tend to have a higher dN and ω than do proteins lacking a predicted coiled-coil (n = 1824 proteins with a predicted coiled-coil, 6685 predicted to lack a coiled-coil; *dN*: $P = 1.6\text{x}10^{-4}$; ω: $P = 0.0035$). This difference in the rate of amino acid evolution does not appear to be due to differences in levels of positive selection, as we find no differences in the proportions of predicted coiled-coil and non-coiled-coil proteins inferred to have experienced positive selection (Table 3.2; $P = 0.147$). Thus, consistent with our prediction, putative coiled-coil proteins appear to be under less amino acid constraint than do proteins lacking predicted coiled-coils.

Discussion and Conclusions

Rapidly evolving proteins, such as the Drosophila seminal fluid prohormone ovulin, represent good empirical examples for studying protein evolvability. Such molecules may help to determine what structural and physicochemical properties of a

Table 3.1 Genome-wide contributors to *dN* and ω. Contributions to *dN* and ω were estimated using multiple linear regression. Factors contributing significantly to either *dN* or ω were chosen according to Larracuente et al. (2008), and presence/absence of a coiled-coil was predicted using PairCoil (McDonnell et al. 2006). 1824 genes encoding predicted coiled-coil proteins and 6685 encoding predicted non-coiled-coil proteins were used for this analysis.

	dN		ω	
Parameter	β	*P*-value	β	*P*-value
Coiled-coil present	0.048	1.6×10^{-4}	0.036	0.0035
Log10(protein length)	0.342	$<2 \times 10^{-16}$	0.217	$<2 \times 10^{-16}$
Specificity	0.473	$<2 \times 10^{-16}$	0.482	$<2 \times 10^{-16}$
Log10(max. expression)	-0.047	1.6×10^{-7}	-0.031	2.3×10^{-4}
Log10(# introns + 1)	-0.495	$<2 \times 10^{-16}$	-0.280	$<2 \times 10^{-16}$
Intron length	-7.5×10^{-6}	1.3×10^{-8}	-2.4×10^{-6}	0.056
Recombination rate	-0.013	0.0011	-0.01	0.0072

Table 3.2 Genome-wide contributors to positive selection. Contributions to positive selection were estimated by logistic regression. Positive selection was inferred using the M7 vs. M8 comparison in PAML, at a 10% FDR.

	dN	
Parameter	β	*P*-value
Coiled-coil present	0.129	0.147
Log10(protein length)	1.41	$<2x10^{-16}$
Specificity	0.769	$<9.8x10^{-11}$
Log10(max. expression)	0.151	0.021
Log10(# introns + 1)	-0.384	0.009
Intron length	$-6.3x10^{-6}$	0.528
Recombination rate	-0.013	0.772

protein contribute to its ability to tolerate variation. Here, we have shown that at least one coiled-coil domain and a tyrosine motif are necessary for the self-interaction of ovulin, a rapidly evolving egg-laying hormone. Both self-interaction motifs are highly conserved between species, and, correspondingly, we have shown that some aspects of tertiary structure are maintained in the face of substantial sequence divergence. These results indicate considerable robustness of tertiary structure to sequence variation. A few key motifs are required for dimerization, but, crucially, the primary sequence requirements for these motifs are minimal. Thus, we propose that the constraints imposed by ovulin's tertiary structure are limited and local to specific domains, such that variation in other parts of the protein does not prevent dimer formation or destabilize the dimer.

Given our hypothesis that ovulin's coiled-coil could contribute to its evolvability, we investigated whether coiled-coils might contribute to evolvability among a broader group of proteins (roughly 60% of those predicted for *D. melanogaster*). Across six species of the genus Drosophila, we found that proteins with predicted coiled-coil motifs tend to have a higher rate of amino acid substitution than do proteins lacking predicted coiled-coil motifs. This difference in substitution rate appears to be due to lower levels of constraint on coiled-coil proteins, as we find no evidence for differences in levels of positive selection between these two groups. Thus, the available evidence is consistent with the hypothesis that coiled-coil motifs, perhaps by virtue of having minimal sequence requirements, reduce overall levels of constraint on a protein. Examination of the evolution rates of proteins bearing other interaction domains with differing degrees of constraint should help to confirm our hypothesis.

Some authors have argued that the evolvability of various biological systems (e.g., the eukaryotic lineage) has itself been selected (Kirschner and Gerhart 1998).

This view has been widely criticized, for example because it seems to require clade-level selection and/or anticipation of future selective environments (e.g., Lynch 2007; Pigliucci 2008). While we have argued that ovulin's coiled-coil structure contributes to its evolvability in virtue of imposing few constraints on its primary sequence, and that coiled-coils may do so more generally, we do not claim that coiled-coils have been selected for this purpose. That is, we do not claim that ovulin's coiled-coil was favored over another dimerization domain *because* it promotes evolvability. We suggest that it is more likely that a highly evolvable coiled-coil protein – ovulin – fortuitously came to be expressed in a milieu subject to strong selective forces, and was thus simply in the right place at the right time.

Regardless of whether protein evolvability is under selection, or is typically a by-product of a protein's functional relationships and structure, the factors contributing to evolvability are of broad interest. Differences in a protein's constraint and robustness to new mutations may be important for understanding why some proteins in similar roles have drastically different evolutionary histories. Moreover, knowledge of attributes contributing to evolvability may inform work on protein engineering, as highly evolvable folds might be more easily adapted to different functions (e.g., Bloom et al. 2005a).

REFERENCES

Aguadé, M. 1998. Different forces drive the evolution of the Acp26Aa and Acp26Ab accessory gland genes in the *Drosophila melanogaster* species complex. Genetics **150**:1079-1089.

Aguadé, M., N. Miyashita, and C. H. Langley. 1992. Polymorphism and divergence in the Mst26A male accessory gland gene region in Drosophila. Genetics **132**:755-770.

Bloom, J. D., S. T. Labthavikul, C. R. Otey, and F. H. Arnold. 2006. Protein stability promotes evolvability. Proc Natl Acad Sci U S A **103**:5869-5874.

Bloom, J. D., M. M. Meyer, P. Meinhold, C. R. Otey, D. MacMillan, and F. H. Arnold. 2005a. Evolving strategies for enzyme engineering. Curr Opin Struct Biol **15**:447-452.

Bloom, J. D., J. J. Silberg, C. O. Wilke, D. A. Drummond, C. Adami, and F. H. Arnold. 2005b. Thermodynamic prediction of protein neutrality. Proc Natl Acad Sci U S A **102**:606-611.

Clark, A. G. et al. 2007. Evolution of genes and genomes on the Drosophila phylogeny. Nature **450**:203-218.

DeLange, R. J., D. M. Fambrough, E. L. Smith, and J. Bonner. 1969. Calf and pea histone IV. 3. Complete amino acid sequence of pea seedling histone IV; comparison with the homologous calf thymus histone. J Biol Chem **244**:5669-5679.

Drummond, D. A., A. Raval, and C. O. Wilke. 2006. A single determinant dominates the rate of yeast protein evolution. Mol Biol Evol **23**:327-337.

Heifetz, Y., L. N. Vandenberg, H. I. Cohn, and M. F. Wolfner. 2005. Two cleavage products of the Drosophila accessory gland protein ovulin can independently induce ovulation. Proc Natl Acad Sci U S A **102**:743-748.

Herndon, L. A., and M. F. Wolfner. 1995. A Drosophila seminal fluid protein, Acp26Aa, stimulates egg laying in females for 1 day after mating. Proc Natl Acad Sci U S A **92**:10114-10118.

Hughes, A. L., and M. Nei. 1988. Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. Nature **335**:167-170.

Kirschner, M., and J. Gerhart. 1998. Evolvability. Proc Natl Acad Sci U S A **95**:8420-8427.

Larracuente, A. M., T. B. Sackton, A. J. Greenberg, A. Wong, N. D. Singh, D. Sturgill, Y. Zhang, B. Oliver, and A. G. Clark. 2008. Evolution of protein-coding genes in Drosophila. Trends Genet **24**:114-123.

Lucey, M. J., D. Chen, J. Lopez-Garcia, S. M. Hart, F. Phoenix, R. Al-Jehani, J. P. Alao, R. White, K. B. Kindle, R. Losson, P. Chambon, M. G. Parker, P. Schar, D. M. Heery, L. Buluwela, and S. Ali. 2005. T:G mismatch-specific thymine-DNA glycosylase (TDG) as a coregulator of transcription interacts with SRC1 family members through a novel tyrosine repeat motif. Nucleic Acids Res **33**:6393-6404.

Lynch, M. 2007. The frailty of adaptive hypotheses for the origins of organismal complexity. Proc Natl Acad Sci U S A **104 Suppl 1**:8597-8604.

McDonnell, A. V., T. Jiang, A. E. Keating, and B. Berger. 2006. Paircoil2: improved prediction of coiled coils from sequence. Bioinformatics **22**:356-358.

Monsma, S. A., H. A. Harada, and M. F. Wolfner. 1990. Synthesis of two Drosophila male accessory gland proteins and their fate after transfer to the female during mating. Dev Biol **142**:465-475.

Monsma, S. A., and M. F. Wolfner. 1988. Structure and expression of a Drosophila male accessory gland gene whose product resembles a peptide pheromone precursor. Genes Dev **2**:1063-1073.

Park, M., and M. F. Wolfner. 1995. Male and female cooperate in the prohormone-like processing of a *Drosophila melanogaster* seminal fluid protein. Dev Biol **171**:694-702.

Pigliucci, M. 2008. Is evolvability evolvable? Nat Rev Genet **9**:75-82.

Ravi Ram, K., L. K. Sirot, and M. F. Wolfner. 2006. Predicted seminal astacin-like protease is required for processing of reproductive proteins in *Drosophila melanogaster*. Proc Natl Acad Sci U S A **103**:18674-18679.

Ravi Ram, K., and M. F. Wolfner. 2007. Seminal influences: Drosophila Acps and the molecular interplay between males and females during reproduction. Integrative and Comparative Biology. **47**: 427-445.

Simmerman, H. K., Y. M. Kobayashi, J. M. Autry, and L. R. Jones. 1996. A leucine zipper stabilizes the pentameric membrane domain of phospholamban and forms a coiled-coil pore structure. J Biol Chem **271**:5941-5946.

Sniegowski, P. D., and H. A. Murphy. 2006. Evolvability. Curr Biol **16**:R831-834.

Swanson, W. J., and V. D. Vacquier. 2002. The rapid evolution of reproductive proteins. Nat Rev Genet **3**:137-144.

Team, R. D. C. 2008. R: A language and environment for statistical computing. R Foundation for statistical computing, Vienna.

Tsaur, S. C., C. T. Ting, and C. I. Wu. 1998. Positive selection driving the evolution of a gene of male reproduction, Acp26Aa, of Drosophila: II. Divergence versus polymorphism. Mol Biol Evol **15**:1040-1046.

Tsaur, S. C., C. T. Ting, and C. I. Wu. 2001. Sex in Drosophila mauritiana: a very high level of amino acid polymorphism in a male reproductive protein gene, Acp26Aa. Mol Biol Evol **18**:22-26.

Tsaur, S. C., and C. I. Wu. 1997. Positive selection and the molecular evolution of a gene of male reproduction, Acp26Aa of Drosophila. Mol Biol Evol **14**:544-549.

Vinson, C., A. Acharya, and E. J. Taparowsky. 2006. Deciphering B-ZIP transcription factor interactions in vitro and in vivo. Biochim Biophys Acta **1759**:4-12.

Wagstaff, B. J., and D. J. Begun. 2005. Comparative genomics of accessory gland protein genes in *Drosophila melanogaster* and *D. pseudoobscura*. Mol Biol Evol **22**:818-832.

Wilke, C. O., J. D. Bloom, D. A. Drummond, and A. Raval. 2005. Predicting the tolerance of proteins to random amino acid substitution. Biophys J **89**:3714-3720.

Wong, A., S. N. Albright, and M. F. Wolfner. 2006. Evidence for structural constraint on ovulin, a rapidly evolving *Drosophila melanogaster* seminal protein. Proc Natl Acad Sci U S A **103**:18644-18649.

Yang, Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol **24**:1586-1591.

CHAPTER 4

EVIDENCE FOR POSITIVE SELECTION ON *DROSOPHILA MELANOGASTER* SEMINAL FLUID PROTEASE HOMOLOGS[1]

Introduction

Evolutionary biologists have long noted that morphological and behavioral traits involved in mating and reproduction diverge rapidly between species (e.g., Darwin 1871; Eberhard 1996). Recently, such observations have been extended to the molecular level, where studies in diverse taxa have found evidence for positive selection on numerous genes involved in reproduction (reviewed in Clark, Aagaard, and Swanson 2006; Panhuis, Clark, and Swanson 2006). Amongst reproductive molecules of animals with internal reproduction, proteins present in the male seminal fluid are of particular interest due to their influences on post-mating behavior and physiology (reviewed in Wolfner 2002; Gillott 2003; Chapman and Davies 2004; Wolfner, Heifetz, and Applebaum 2005; Clark, Aagaard, and Swanson 2006; Poiani 2006), and hence their importance to reproductive success.

In *D. melanogaster*, seminal fluid includes over 100 proteins produced by the male's accessory glands (hereafter Acps for accessory gland proteins), as well as proteins expressed in the ejaculatory bulb and the ejaculatory duct (reviewed in Wolfner 2002; Chapman and Davies 2004; Wolfner, Heifetz, and Applebaum 2005; Ravi Ram and Wolfner 2007). Studies using males that lack specific Acps, or that make no Acps at all, have demonstrated roles for these proteins in diverse post-mating

[1]This chapter was published previously as: Wong A, Turchin MC, Wolfner MF, Aquadro CF. 2008. Evidence for positive selection on *Drosophila melanogaster* seminal fluid protease homologs. *Molecular Biology and Evolution* 25: 497-506. Michael Turchin amplified and sequenced one locus for this study (CG11864) under my supervision. I performed the remaining sequencing and analyses, and co-wrote the manuscript with CFA and MFW. Copyright permissions for theses are automatically granted by the journal.

processes, e.g., sperm storage, egg-production and egg-laying, increased female mortality, decreased female receptivity, and increased feeding (Kalb, DiBenedetto, and Wolfner 1993; Chapman et al. 1995; Tram and Wolfner 1998; Heifetz et al. 2000; Xue and Noll 2000; Carvalho et al. 2006; Chapman et al. 2003; Kubli 2003; Liu and Kubli 2003; Adams and Wolfner 2007). Moreover, mutational and knockdown analyses have ascribed specific roles to ~8 individual Acps (Aigaki et al. 1991; Herndon and Wolfner 1995; Neubaum and Wolfner 1999; Heifetz et al. 2000; Liu and Kubli 2003; Wigby and Chapman 2005; Ravi Ram, Sirot, and Wolfner 2006; Ravi Ram and Wolfner in press). Males lacking the prohormone ovulin (Acp26Aa), for example, induce less ovulation in their mates during the first 24 hours post-mating than do wild-type males (Herndon and Wolfner 1995; Heifetz et al. 2000), and the large glycoprotein Acp36DE is necessary for normal levels of sperm storage (Neubaum and Wolfner 1999; Bloch Qazi and Wolfner 2003). In addition to such knockout and knockdown approaches, association studies have suggested roles for several Acps (e.g., Acp29AB) in sperm competition (Clark et al. 1995; Fiumera, Dumont, and Clark 2005; Fiumera, Dumont, and Clark 2007).

Several studies have shown that genes encoding Drosophila Acps evolve differently on average than do other classes of genes. Acp genes show an elevated average level of amino acid divergence (d_N) compared to non-reproductive genes in comparisons between the closely related species *D. melanogaster* and *D. simulans*, despite similar synonymous site divergence (d_S) (Swanson et al. 2001; Mueller et al. 2005). This observation holds true across a broader phylogenetic range as well – across the genomes of six fully sequenced species in the *melanogaster* species group, Haerty et al. (in press) found that mean d_N/d_S for 25 genes encoding seminal fluid proteins (including several Acps) is significantly higher than that of ~8500 genes not encoding seminal fluid proteins. Acp genes also tend to have lower levels of codon

bias than do non-Acps (Begun et al. 2000; Mueller et al. 2005), and show an over-dispersal of amino acid substitutions within the *D. simulans* species complex (Kern, Jones, and Begun 2004).

Positive selection appears to account for at least some of the unusual patterns of Acp molecular evolution. For several Acp genes, patterns of polymorphism and divergence are consistent with positive selection in comparisons between *D. melanogaster* and *D. simulans* (Aguadé, Miyashita, and Langley 1992; Tsaur and Wu 1997; Tsaur, Ting, and Wu 1998; Aguadé 1999; Begun et al. 2000; Holloway and Begun 2004). Moreover, reduced variation at the Acp loci Acp36DE (Begun et al. 2000) and Lectin29Ca (Holloway and Begun 2004) in non-African populations of *D. melanogaster* has been interpreted as evidence for recent selective sweeps at these loci. Moreover, recent divergence analyses find evidence for positive selection on a subset of codons for each of 9 additional Acp loci within the *melanogaster* species group (Haerty et al. in press). Finally, polymorphism and divergence analyses on several putative Acp-encoding loci also provide substantial evidence for positive selection in *D. pseudoobscura* (Stevison, Counterman, and Noor 2004; Wagstaff and Begun 2005a; Schully and Hellberg 2006) and in two cactophilic species of *Drosophila* (Wagstaff and Begun 2005b). Nonetheless, a relatively small proportion of Acp loci have been examined in most previous studies, making generalizations about selective regimes difficult.

Several hypotheses have been forwarded to explain the rapid amino acid evolution of, and increased incidence of positive selection amongst, *Drosophila* Acps and seminal fluid proteins more generally (reviewed in Swanson and Vacquier 2002; Clark, Aagaard, and Swanson 2006; Panhuis, Clark, and Swanson 2006). First, male-female and male-male interactions may underlie rapid Acp evolution. Sexual conflict, sperm competition, cryptic female choice, and other forms of sexual selection may

exert strong selective pressures on some Acps, given the roles of these proteins in clearly relevant physiological processes (e.g., sperm storage, post-mating lifespan reduction). Second, host-pathogen interactions may account for some rapid Acp evolution. *Drosophila* seminal fluid contains several components with anti-bacterial activity (Samakovlis et al. 1991; Lung, Kuo, and Wolfner 2001; Mueller, Page, and Wolfner 2007), which may help to prevent infection during mating. Thus, Acps with roles in immunity may evolve rapidly as a consequence of a host-pathogen arms race.

In this study, we focus on the molecular evolution and molecular population genetics of five predicted proteases and protease homologs (i.e., proteins resembling proteases in sequence and structure, but with one or more catalytic site mutations; Ross et al. 2003) expressed in the *D. melanogaster* male accessory gland. Three lines of evidence suggest that reproductive tract proteases and protease homologs may be subject to strong selection. First, proteolysis regulators (used here to refer to proteases and their inhibitors, as well as protease homologs) are likely to mediate male-female interactions during mating. Previous work has suggested roles for both male and female derived factors in the processing of at least one Acp prohormone (ovulin; Park and Wolfner 1995; Ravi Ram, Sirot, and Wolfner 2006), and numerous proteolysis regulators are present in both male seminal fluid and in the female reproductive tract (Swanson et al. 2001; Mueller et al. 2004; Swanson et al. 2004; Mack et al. 2006; Kelleher, Swanson, and Markow 2007; Lawniczak and Begun 2007). As such, interactions between male- and female-derived proteolysis regulators may be foci for sexual selection. Consistent with this hypothesis, several proteases expressed in the Drosophila female reproductive tract show evidence of positive selection (Swanson et al. 2004; Panhuis and Swanson 2006; Kelleher, Swanson, and Markow 2007; Lawniczak and Begun 2007), and one Acp protease inhibitor (Acp76A) out of two that

have been examined shows evidence for positive selection along the *D. simulans* lineage (Begun et al. 2000; Kern, Jones, and Begun 2004).

Second, proteases, protease homologs, and protease inhibitors are associated with fertility effects in several species, including Drosophila (Ravi Ram and Wolfner, in press) and mice (Mbikay et al. 1997; Murer et al. 2001; Carpentier et al. 2004; Nie et al. 2005; Uhrin et al. 2007), again raising the possibility that proteolysis regulators are subject to sexual selection. Moreover, several predicted protease inhibitors present in male seminal fluid are toxic upon ectopic expression, and may therefore contribute to the cost of mating (Lung et al. 2002; Mueller, Page, and Wolfner 2007). Third, proteolytic cascades play important roles in immunity and defense in many organisms (Ligoxygakis et al. 2002a; Ligoxygakis et al. 2002b; Sim and Tsiftsoglou 2004), and thus may experience selection pressure from pathogens.

In this study, we report results from molecular population genetic surveys and divergence analyses of five Acp genes encoding predicted proteases or protease homologs (see Table 4.1 for gene symbols, gene ontologies, and coding sequence lengths). These genes encode five out of the six protease/protease homologs reported in an EST screen of the male accessory gland (Swanson et al. 2001), although more genes encoding predicted proteases or protease homologs with accessory gland biased expression have since been identified (Chintapalli, Wang, and Dow 2007; Ravi Ram and Wolfner 2007). Two of these five genes, CG11864 and CG6168, are each predicted to encode a metalloprotease (Mueller et al. 2004), i.e., a protease with a metal ion at its active site. Previous studies have assigned potential physiological roles to both metalloproteases: CG11864 is essential for cleavage of at least two other Acps, Ovulin/Acp26Aa and Acp36DE (Ravi Ram, Sirot, and Wolfner 2006), and ectopic expression of CG6168 in a virgin female increases her ability to clear a bacterial infection (Mueller, Page, and Wolfner 2007).

Table 4.1 Genes examined in chapter 4. A serine protease homolog resembles a catalytically active serine protease, but bears one or more active site mutations, suggesting that catalytic function is likely absent. Such molecules have been proposed to regulate proteolytic cascades (Kwon et al. 2000; Lee et al. 2002; Asgari et al. 2003; Jiang et al. 2003a; Jiang et al. 2003b; Yu et al. 2003; Gupta, Wang, and Jiang 2005).

Gene	Codons	Gene Ontology
CG6069	283	Serine protease homolog
CG6168	314	Metalloprotease
CG9997	330	Serine protease homolog
CG11664	209	Serine protease homolog
CG11864	251	Metalloprotease

The other three genes examined here, CG6069, CG9997, and CG11664, are predicted to encode serine protease homologs – although they are predicted to resemble serine proteases in overall structure, mutations at one or more of the three canonical active site residues likely render them non-catalytic (Mueller et al. 2004). Of these three genes, a function has thus far only been assigned to one: RNAi knockdown of CG9997 suggests that it is essential for normal sperm usage (Ravi Ram and Wolfner, in press). We note that the biochemical and physiological roles of protease homologs are currently not well understood. Although these proteins likely lack catalytic activity, several studies suggest roles for protease homologs in regulating the activity of catalytically active proteases, either as co-factors or as competitive inhibitors (Kwon et al. 2000; Lee et al. 2002; Asgari et al. 2003; Jiang et al. 2003a; Jiang et al. 2003b; Gupta, Wang, and Jiang 2005). Such roles would make proteases and protease homologs alike subject to the evolutionary pressures just described.

Using molecular population genetic surveys of African population samples of *D. melanogaster* (Pool and Aquadro 2006), we find evidence for strong directional selection at two loci out of five examined, CG6069 and CG9997. At a deeper evolutionary time scale, we also find evidence for recurrent positive selection on a subset of codons in CG6069. These findings, along with previous studies on reproductive tract proteolysis regulators in male (Kern, Jones, and Begun 2004) and female *Drosophila* (Swanson et al. 2004; Panhuis and Swanson 2006; Kelleher, Swanson, and Markow 2007; Lawniczak and Begun 2007), support the hypothesis that interactions between males and females drive the rapid evolution of some reproductive genes.

Materials and Methods

Drosophila Strains, DNA Sequences, and Sequence Alignment

For polymorphism-based analyses, we used chromosome extraction lines derived from African populations of *D. melanogaster* (Uganda for chromosomes 2 and 3, Zimbabwe for the X; population samples are described in Pool and Aquadro 2006), with *D. simulans* as an outgroup (*D. simulans* sequences reported in Mueller et al. 2005). Sample sizes for each gene are given in Table 4.2. For divergence analyses, we used 6 species in the *D. melanogaster* subgroup. Sequences from *D. melanogaster*, *D. simulans*, and *D. yakuba* are from Mueller et al. (2005). Additional sequences were collected from *D. teissieri* (Tucson *Drosophila* stock center *D. teissieri* 257.0), *D. erecta* (S-18; originally from the Ashburner laboratory), and *D. santomea* (strain CAR1566-8, kindly donated by Peter Andolfatto).

DNA extractions were performed using the Puregene DNA purification kit (Gentra Systems), and genes were amplified by polymerase chain reaction (PCR); primer sequences and conditions are available upon request. Sequencing was carried out on an ABI 3730 automated sequencer using BigDye chemistry (Applied Biosystems). Sequence alignments were carried out using the ClustalW algorithm as implemented in MegAlign (DNASTAR, Inc.) on protein sequences. Sequences have been deposited in GenBank under accession numbers EU325840-EU328631. Introns in CG11864 were identified in other species using the *D. melanogaster* annotation as a guide; the other genes contain no introns in their coding regions. Consensus sites for intron start (AG) and stop (GT) were conserved across all species.

Analysis of polymorphism

Summary statistics (θ, π, and divergence) for each gene were calculated using DnaSP4.1 (Rozas et al. 2003). McDonald-Kreitman tests (McDonald and Kreitman

Table 4.2 Summary statistics for five protease/protease homolog encoding genes. Population summary statistics are for African populations of *D. melanogaster*. n: alleles sequenced, S: number of segregating sites. d_N and d_S were calculated using *D. melanogaster* and *D. simulans*. D: Tajima's D statistic (Tajima 1989). H: Fay and Wu's H statistic (Fay and Wu 2000). No locus rejects neutrality using D or H.

Gene	**n**	**Length** (bp)	**S**	**theta**	π_{Tot}	π_{Syn}	π_{Non}	d_S	d_N	**D**	**H**
CG6069	20	945	8	0.0025	0.0018	0.0060	0.0002	0.136	0.0161	-0.97	0.85
CG6168	18	905	80	0.0275	0.0228	0.0579	0.0128	0.178	0.0360	-0.56	0.73
CG9997	12	912	17	0.0062	0.0066	0.0228	0.0017	0.099	0.0251	0.26	2.24
CG11664	12	641	12	0.0041	0.0038	0.0136	0.0007	0.194	0.0247	-0.29	0.55
CG11864	12	694	11	0.0053	0.0053	0.0099	0.0029	0.119	0.033	-0.52	-2.27

1991), as well as Tajima's test (Tajima 1989) and Fay and Wu's H test (Fay and Wu 2000), were also performed using DnaSP. *D. simulans* was used for all analyses requiring an outgroup. In order to calculate statistical significance for Tajima's D and Fay and Wu's H, we used the coalescent simulator with recombination implemented in DnaSP. The population recombination rate $R = 4N_e rm$ was estimated using $N_e = 1 \times 10^6$ (Kreitman 1983), where m is the size in base pairs of the gene under consideration (Table 4), and with per base pair recombination rate estimates *r* obtained from (Hey and Kliman 2002). Estimated values of *R* were 100.65 for CG6069, 138.82 for CG6168, 108.84 for CG9997, 9.83 for CG11664, and 12.29 for CG11864.

Hudson-Kreitman-Aguadé (HKA) tests were performed using the maximum-likelihood method of Wright and Charlesworth (2004) (Hudson, Kreitman, and Aguadé 1987; Wright and Charlesworth 2004). This method uses loci specified *a priori* to generate a null model of sequence evolution, and assesses the fit of one or more loci of interest to that null model. The parameter *k* measures the decrease or increase of polymorphism relative to divergence, with the neutral expectation that $k = 1$. We used four X-linked non-coding loci reported in Pool and Aquadro (2006) as representative 'neutral' loci, and tested each protease or protease homolog encoding gene individually.

Divergence based analyses

Inferences of positive selection using comparisons between the number or rates of nonsynonymous and synonymous substitutions can be misled if the latter varies across a sequence. If, for example, some sites have a particularly low rate of synonymous substitution, $\omega > 1$ may be inferred even in the absence of positive selection, under the assumption of a single synonymous rate. As such, we used maximum-likelihood methods implemented in HyPhy (Pond, Frost, and Muse 2005)

to test for heterogeneity in the rate of synonymous substitution (d_S) at different codons in each multiple sequence alignment. The null model assumes no variation in d_S, but allows variation in the rate of nonsynonymous substitution (d_N) in the form of three discrete rate classes (Pond and Muse 2005). The alternative model allows variation in d_S, with two discrete rate classes. The two models can be compared using either (a) A likelihood ratio test, with twice the difference in –lnL between models following a χ_4^2 distribution, or (b) Akaike Information Criterion (AIC) scores. Rejection of the null model provides evidence for variation in d_S at different codons.

Sequence alignments were analyzed for evidence of positive selection in the form of an elevated rate of non-synonymous substitution compared to the rate of synonymous substitution (ω) using PAML (Yang 1997; Yang et al. 2000). Two model comparisons were performed. In the first comparison, the null model M1a allows for two classes of sites: one with $0 < \omega < 1$, and one with $\omega = 1$. The alternative model M2a adds a third site class with $\omega > 1$. In the second comparison, the null model M8A uses a beta distribution to describe sites with $0 < \omega < 1$, with an extra category of sites with $\omega = 1$. The alternative model M8 allows the extra category to undergo positive selection, i.e. requires $\omega > 1$. For both comparisons, the null and alternative models can be compared via likelihood ratio test (LRT), with the difference in log likelihoods (δ) following a χ_2^2 distribution (M1a vs. M2a) or a χ_1^2 distribution (M8A vs. M8).

In order to evaluate the fit of δ to the appropriate χ^2 distribution, and as an independent estimate of the *P* value for each model comparison, we also implemented a parametric bootstrap. Parameter estimates from M1a were used to generate 250 simulated datasets using evolverNSsites (Anisimova, Bielawski, and Yang 2001), which were then analyzed under M1a, M2a, M8A, and M8. Values of δ from the simulated neutral datasets were then calculated and used to obtain the probability of obtaining the observed value of δ under the null hypothesis.

For both HyPhy and PAML analyses, a single tree – ((*D. melanogaster*, *D. simulans*), (*D. teissieri*, (*D. yakuba*, *D. santomea*)), *D. erecta*) – was assumed, following current understanding (e.g., Ko, David, and Akashi 2003; LaChaise et al. 2000; Pollard et al. 2006; Wong et al. 2007). We note that several authors have suggested that lineage sorting has occurred in the common ancestor of the *melanogaster* subgroup, and that this may introduce inferential problems (Pollard et al. 2006; Wong et al. 2007). Here, however, all analyses were conducted on an unrooted tree, with no species basal to the *melanogaster* subgroup; as such, our analyses should not be affected by lineage sorting.

Results

Various modes of adaptation can leave different signatures in sequence data, such that different kinds of sample are suitable for their detection. For example, repeated episodes of positive selection on a few codons of a coding sequence can be inferred using multi-species divergence data, while polymorphism data from a single population is suited to the detection of a recent selective sweep or ongoing balancing selection. We have collected both divergence and polymorphism data in order to gain a comprehensive view of the patterns of molecular evolution at five putative protease-encoding Acp genes.

Polymorphism analyses

We collected polymorphism data for each gene from African populations of *D. melanogaster*. We used an African population rather than a North American, European, or Asian population in order to avoid, as best as possible, inferential problems stemming from non-equilibrium demographic histories (e.g., Jensen et al. 2005; Thornton et al. 2007). 12-20 alleles were sequenced for each gene; sample sizes

and summary statistics are given in Table 4.2. Neither Tajima's D (Tajima 1989) nor Fay and Wu's H (Fay and Wu 2000) deviates from the neutral expectation for any gene (Table 4.2).

We used two additional tests of neutrality to assess the fit of the polymorphism data to the standard neutral model. The first, the McDonald-Kreitman (MK) test (McDonald and Kreitman 1991), tests the neutral prediction that the ratio of non-synonymous to synonymous substitutions between species should equal the ratio of non-synonymous to synonymous polymorphisms within species (Table 4.3). For four genes, CG6069, CG6168, CG11664, and CG11864, we fail to find any deviation from the null hypothesis. However, for CG9997, the null hypothesis of equal ratios of non-synonymous to synonymous changes within and between species is rejected (P = 0.008). Rejection of the null hypotheses could in theory result from deviations from the neutral expectation in any cell of the MK table; we suggest that an excess of non-synonymous fixations is the most likely explanation. Nonsynonymous divergence is high at CG9997 (d_N = 0.025 for the *melanogaster*/*simulans* comparison; Table 4.2), relative to an average d_N of 0.0124 (95% CI: 0.0121 – 0.0128) compiled from ~8500 genes (data from Larracuente et al., submitted). This does not appear to be the result of a high mutation rate at CG9997, since synonymous divergence (d_s = 0.099) is slightly lower than average (0.128; 95% CI: 0.126 – 0.129). In addition, levels of polymorphism at CG9997 do not appear to differ substantially from average. Thus, it is likely that an excess of nonsynonymous substitution due to positive selection in the lineages leading to *D. melanogaster* and/or *D. simulans* accounts for this result.

We also used the HKA-test (Hudson, Kreitman, and Aguadé 1987) to assess the neutral prediction that the ratio of polymorphism to divergence should be the same for different loci (Table 4.4). This test is particularly useful for detecting a deficit or an excess of polymorphism due to recent directional selection or balancing selection,

Table 4.3 McDonald-Kreitman tests for five protease/protease homolog genes. *P*-values were obtained using a 2-tailed Fisher's exact test.

			Polymorphic		Fixed		
Gene	Population	n	Silent	Replacement	Silent	Replacement	Prob.
CG6069	Uganda	20	5	1	25	10	0.665
CG6168	Uganda	18	45	36	25	16	0.699
CG9997	Uganda	12	14	3	16	23	0.008
CG11664	Zimbabwe	12	6	2	18	12	0.684
CG11864	Uganda	12	5	5	15	11	0.722

Table 4.4 Maximum-likelihood HKA tests for 5 protease/protease homolog genes. Silent S: Synonymous segregating sites, Silent Divergence: Synonymous divergence between *D. melanogaster* and *D. simulans*, 2 * ΔlnL: Twice the difference in log-likelihood between the null and selection models, *P*: P-value obtained from a χ^2 test (df = 1), k: Estimated ratio of variation at the given locus to the neutral expectation.

Gene	Silent S	Silent Divergence	2 * ΔlnL	*P*	k
CG6069	5	27	5.83	0.016*	0.22
CG6168	45	35	0.83	0.36	1.57
CG9997	14	24	0.04	0.85	0.8
CG11664	8	20	0.26	0.61	0.71
CG11864	5	18.5	1.79	0.17	0.41

respectively. We used four X-linked non-coding loci (named after their cytological locations: 4F2, 8A4, 11A5, and 12F1) reported in Pool and Aquadro (2006) as representative 'neutral' loci, and tested each protease- or protease homolog-encoding gene against the neutral prediction, using the maximum-likelihood HKA test of Wright and Charlesworth (2004). We found that four protease/protease homolog genes conformed to the neutral prediction, with one protease homolog gene, CG6069, rejecting neutrality ($P = 0.016$). The latter rejection of neutrality could be the result of either elevated silent site divergence or a deficit of silent polymorphism. Given that d_S is about average for CG6069 (0.139 for CG6069 vs. 0.128 genome-wide), while very few polymorphisms were observed ($\theta_S = 0.00694$), we suggest that CG6069 is depauperate for variation, consistent with the action of recent selection at or near this locus. The rate of recombination in this region of the genome is moderate ($r = 2.6$; Hey and Kliman 2002), consistent with CG6069 (rather than a linked locus) being the target of selection. Tests of the frequency spectrum (Tajima's D, Fay and Wu's H) do not reject neutrality; we suspect that low variation (perhaps due to very recent selection) reduces the power of these tests.

CG6168 presents an interesting case. Polymorphism at this gene is extremely high (Table 4.2), with 80 segregating sites in the coding region and $\pi_S = 0.0579$ (vs. an average of ~0.029 genome-wide; Andolfatto 2005), yet the HKA-test does not reject neutrality. Similarly, tests of neutrality based on site-frequency spectra do not find deviations from the neutral expectation, either using all polymorphisms (Table 4.2) or synonymous and non-synonymous polymorphisms separately (Table 4.5). Polymorphism is also very high in a population sample collected from Pennsylvania, despite a recent bottleneck for non-African populations of *D. melanogaster* (Anthony Fiumera, personal communication). High silent site divergence at CG6168 ($d_S = 0.178$; Table 4.2) may account for the failure to reject neutrality using the HKA-test.

Table 4.5 Comparisons of frequency spectra for synonymous and nonsynonymous polymorphisms at CG6168

	Tajima's D	Fu and Li's D	Fay and Wu's H
Synonymous	-0.500	-0.851	-1.935
Nonsynonymous	-0.756	-1.044	-0.418

There is no evidence that selection drives high synonymous site divergence, as patterns of unpreferred and preferred differences within and between species are not significantly different using Akashi's (1999) fddMWU test (Akashi 1999). We suspect that balancing selection may operate to maintain high levels of polymorphism at this locus, but more data will be required to rigorously evaluate this hypothesis.

Divergence analyses

Variation in d_S within a gene can mislead commonly used individual locus divergence-based tests for positive selection (Pond and Muse 2005). Using model comparisons implemented in HyPhy (Pond, Frost, and Muse 2005), we fail to find evidence of variation in d_S at any gene examined in this study. For all five genes, the data do not fit a model incorporating variation in d_S significantly better than they fit a null model with no such variation (Table 4.6). Although failure to reject the null hypothesis does not warrant its acceptance, this result suggests that use of models that assume a single synonymous substitution rate to infer positive selection should not be misled by variation in d_S.

We therefore used PAML, which assumes a single value of d_S for each gene, to infer the action of recurrent positive selection on individual codons (Table 4.7). We find strong evidence for positive selection on one gene, CG6069. Using both the M2a vs. M1a and the M8 vs. M8A comparisons, the data for CG6069 fit the alternative (selection) model significantly better than they do the null model. 4-5% of codons are estimated to belong to the selected class, with $\omega = 3.44$ under M8 ($\omega = 3.98$ under M2a). Since the predicted three dimensional structure of the protein encoded by CG6069 was previously modeled (Mueller et al. 2004), we could locate the putative positively selected residues on its predicted structure (Figure 4.1). The sites whose mean ω +/- 1 SE is greater than 1 (corresponding to posterior probabilities >0.774 of

Table 4.6 Tests for variation between sites in the rate of synonymous substitution using HyPhy. MG94 x REV Nonsynonymous GDD 3 is a model incorporating variation in the rate of nonsynonymous substitution (3 rate classes), but not in the rate of synonymous substitution. The dual model incorporates variation in both the rate of nonsynonymous substitution (3 rate classes) and the rate of synonymous substitution (2 rate classes). The indicated *P*-value is for the likelihood ratio test between the nonsynonymous and dual models, using the asymptotic distribution of χ^2_4. No tests were significant with $\alpha = 0.05$, indicating no evidence for variation in the rate of synonymous substitution. ΔAIC: Difference in Akaike Information Criterion (AIC) scores between the nonsynonymous and dual models; negative ΔAIC values indicate that the dual GDD 2 x 3 model does not outperform the GDD 3 model, given the extra parameters used by GDD 2 x 3.

Gene	logL		*P*-value	ΔAIC
	MG94 x REV Nonsynonymous GDD 3	MG94 x REV Dual GDD 2 x 3		
CG6069	-2479.46	-2475.53	0.097	-0.16
CG6168	-2582.97	-2580.94	0.398	-3.93
CG9997	-2588.96	-2588.14	0.803	-6.37
CG11664	-1626.42	-1626.29	0.992	-7.74
CG11864	-1943.62	-1943.43	0.989	-7.62

Table 4.7 Tests for positive selection using PAML. -lnL: Negative log-likelihood for the indicated model. *P*-values: The first *P*-value reported is for the likelihood ratio test between the selection (M2a or M8) and neutral (M1a or M8A) models, using the asymptotic distribution of χ_2^2 (M2a vs. M1a) or χ_1^2 (M8 vs. M8A). The second *P*-value was obtained by parametric bootstrapping under the maximum likelihood parameter estimates from model M1a. 250 bootstrap replicates were generated using evolverNSsites. The last two columns give estimated values of ω under the indicated model. Numbers in parentheses indicate the proportion of codons estimated to belong to the selected class, for comparisons where positive selection was inferred.

Gene	-lnL				*P*-value			
	M1a	M2a	M8A	M8	M2a vs. M1a	M8 vs. M8A	ω M2a (prop)	ω M8 (prop)
CG6069	2424.71	2420.67	2424.72	2420.42	0.018*; 0.004*	0.003*; 0.004*	3.98 (0.038)	3.44 (0.055)
CG6168	2328.43	2328.43	2328.32	2328.32	1; 1	1; 1	1	1
CG9997	2348.63	2348.32	2348.64	2348.33	0.738; 0.280	0.432; 0.308	2.9	1.60
CG11664	1574.68	1573.33	1574.69	1573.06	0.06; 0.259	0.048*; 0.071	3.15	3.41
CG11864	1690.86	1690.86	1690.86	1690.86	1; 1	1; 1	1	1

Figure 4.1 Structural model of the predicted protease homolog encoded by CG6069. Sites whose inferred $\omega \pm 1$ standard error is greater than 1 are shown in white. All five selected residues that fall within the modeled domain are predicted to lie on the proteins surface, although none lies in the predicted substrate binding cleft. The model was generated by Mueller et al. (2004).

belonging to the selected class) that fell within the modeled domain (five out of six total) are predicted to be on the protein's surface, although none lies within the predicted substrate binding cleft. For a second gene, CG11664, the M8 vs. M8A comparison is marginally significant using a χ^2-test ($P = 0.048$), with other tests being marginally non-significant.

Use of a parametric bootstrap to evaluate the significance of model comparisons was consistent with the results obtained from likelihood ratio tests (Table 7). In most cases, the LRT and the bootstrap resulted in rejections, or failures to reject, for the same comparisons. The one exception is for the M8A vs. M8 comparison for CG11664, where the LRT result is marginally significant ($P = 0.048$) and the bootstrap result is non-significant ($P = 0.071$).

Since neutrality was rejected for CG6069 using both divergence and polymorphism based tests (Tables 4 and 7), we were interested in determining whether recent selection on this gene has targeted the same residues as those identified as under positive selection by PAML. Of the six codons identified by PAML as having $\omega > 1$, two – ^{190}Ile and ^{268}Ser - appear to have changed along the *melanogaster* species lineage, although the high variability of both codons makes polarization of changes uncertain. These two codons are particularly good candidates for having been recent targets of selection in *D. melanogaster*. An additional six sites, ^{124}Ile, ^{152}Ser, ^{208}Ile, ^{230}Gly, and ^{285}Thr, appear to have fixed along the *melanogaster* lineage but do not have high posterior probabilities of $\omega > 1$.

Discussion

A number of genes encoding seminal fluid proteins show evidence for positive selection in diverse taxa, e.g., *Drosophila* (reviewed in Clark, Aagaard, and Swanson 2006; Panhuis, Clark, and Swanson 2006), crickets (Andres et al. 2006), and primates

(Clark and Swanson 2005). Several explanations have been proposed for the rapid, adaptive evolution of genes encoding seminal fluid proteins, including post-mating male-female or male-male interactions and immune pressures. We hypothesized that some Acp proteases would be targets of adaptive evolution in *D. melanogaster* and its close relatives, given the potential role of proteolysis regulators in mediating male-female interactions, and known or suspected roles for several such proteins in immunity, sperm usage, and proteolytic processing of other rapidly evolving Acps. Using polymorphism-based tests, we find evidence for positive selection on two protease homolog genes out of five genes examined: CG9997 appears to have undergone an excess of amino acid substitutions between *D. melanogaster* and *D. simulans*, while patterns of polymorphism at CG6069 are consistent with a recent selective sweep. Furthermore, between-species analyses suggest that CG6069 has experienced pervasive positive selection on a subset of codons in the *melanogaster* subgroup.

RNAi knockdown studies on CG9997 suggest a role for this gene's product in regulating the release of sperm from storage in females (Ravi Ram and Wolfner, in press). Since sperm storage is potentially involved in cryptic female choice and sperm competition (Eberhard 1996; Simmons 2001), it is likely that sexual selection of some variety underlies the molecular evolution of this gene. However, it should be noted that CG9997's role in other systems potentially subject to strong selection, e.g., the immune response, has not been fully investigated. Ectopic expression of CG9997 in females does not affect systemic clearance of the gram-negative bacterium *S. marcescens* (Mueller, Page, and Wolfner 2007), but its activity against gram-positive bacteria or fungi, or any localized activity in the reproductive tract, has not been examined.

Knockdown and ectopic expression studies have not yet uncovered any potential role for CG6069, the second positively selected gene identified here, in the regulation of post-mating responses (Ravi Ram and Wolfner, in press), seminal fluid toxicity (Mueller, Page, and Wolfner 2007), or in immunity (Mueller, Page, and Wolfner 2007). Moreover, no data currently exist with respect to the localization of CG6069's protein product in the female reproductive tract. As such, it is not currently possible to ascribe this gene's rapid molecular evolution to a particular physiological process.

CG9997 and CG6069, the two genes inferred in this study to have experienced positive selection, are predicted to encode serine protease homologs (SPHs), i.e., their protein products are predicted to resemble serine proteases, but bear mutations in one or more of the three canonical active site residues (Ross et al. 2003; Mueller et al. 2004). As such, these proteins are probably not proteolytically active. However, non-catalytic roles have assigned to, or suggested for, SPHs in several systems. For example, studies on the cleavage of prophenoloxidase (proPO) to phenoloxidase (PO), which is involved in the melanization of pathogens, have suggested a role for SPHs in modulating the activity of proPO activating proteases (PAPs). In tobacco hornworm (Jiang et al. 2003a; Jiang et al. 2003b; Yu et al. 2003; Gupta, Wang, and Jiang 2005) and several beetles (Kwon et al. 2000; Lee et al. 2002), SPHs are required for full proteolytic activity of PAPs. Conversely, a SPH present in the venom of a parasitic wasp is capable of interfering with proPO cleavage, perhaps by competing with host SPHs for binding to PAP and/or proPO (Asgari et al. 2003). If positively selected *Drosophila* SPHs present in the seminal fluid function either as agonists or antagonists of catalytically active proteases, then co-evolution with proteases, protease substrates, inhibitors, or other binding partners may underlie their adaptive evolution. Other documented molecular functions for SPHs include glycoprotein binding (Watorek

2003) and cell adhesion (Huang et al. 2000; Lin et al. 2006); seminal fluid SPHs could also be involved in any of these functions, as a number of other Acps are glycosylated (Monsma and Wolfner 1988; Bertram, Neubaum, and Wolfner 1996; Saudan et al. 2002), and cell adhesion may be important for sperm storage and/or fertilization.

We found no evidence for positive selection on 3 other protease-encoding Acp genes, 2 of which, CG11864 and CG6168, have been ascribed functions using genetic methods. Knockdown of CG11864 shows that this putative metalloprotease is necessary for the proteolytic cleavage of two Acps, the egg-laying prohormone ovulin and the sperm storage protein Acp36DE (Ravi Ram, Sirot, and Wolfner 2006). While both ovulin and Acp36DE appear to have experienced positive selection (Aguadé, Miyashita, and Langley 1992; Begun et al. 2000; Fay and Wu 2000), we found no evidence of a similar history for CG11864. This is not, we suggest, a surprising result: If proteolytic cleavage of ovulin and/or Acp36DE is necessary for some aspect of their functions (although there is no evidence to suggest this to date), then both the cleavage sites and the responsible protease(s) should be well-conserved. We note that other regions of ovulin thought be structurally important are highly conserved between species (Wong, Albright, and Wolfner 2006), and suspect that the same will be true of cleavage sites.

Ectopic expression of the predicted metalloprotease CG6168 in females aids in the clearance of systemic *S. marcescens* infection (Mueller, Page, and Wolfner 2007), suggesting that this protein may participate in immune regulatory cascades. Polymorphism at CG6168 is high, although several tests find no deviations from the neutral expectation. Classic studies attribute extremely high levels of polymorphism at MHC genes to balancing selection arising from a host-pathogen interactions (e.g., Hughes and Nei 1988; McConnell et al. 1988). It is possible that a similar explanation

underlies high polymorphism at CG6168, but further statistical and functional analyses are required.

Conclusions

An understanding of the rapid evolution of an elevated proportion of *Drosophila* Acps, and reproduction-related genes more generally, requires both extensive sequence data and functional characterization. Full genome sequences from multiple species of *Drosophila* have allowed a comprehensive examination of sex- and reproduction-related genes on a deep phylogenetic scale (Haerty et al. in press). Population genetic analyses, however, have been narrower in scope, with most studies focusing on a limited set of genes. We conducted divergence and polymorphism analyses at five male Acp-encoding loci that have not been previously examined at the population level, and found evidence for positive selection at two predicted protease homolog encoding genes. Adaptive evolution of protease, protease homolog, or protease inhibitor genes has now been documented in genes expressed in either the male accessory gland (this study; Kern, Jones, and Begun 2004) or the female reproductive tract (Swanson et al. 2004; Panhuis and Swanson 2006; Kelleher, Swanson, and Markow 2007; Lawniczak and Begun 2007). While definitive interpretation of these results must await functional characterization of positively selected genes (data from females are particularly lacking), the finding of positive selection on both male and female reproductive tract genes suggests that between-sex interactions, rather than simply male-male competition, drives the rapid evolution of some reproductive genes.

REFERENCES

Adams, E. M., and M. F. Wolfner. 2007. Seminal proteins but not sperm induce morphological changes in the *Drosophila melanogaster* female reproductive tract during sperm storage. J Insect Physiol 53:319-331.

Aguadé, M. 1999. Positive selection drives the evolution of the Acp29AB accessory gland protein in Drosophila. Genetics 152:543-551.

Aguadé, M., N. Miyashita, and C. H. Langley. 1992. Polymorphism and divergence in the Mst26A male accessory gland gene region in Drosophila. Genetics 132:755-770.

Aigaki, T., I. Fleischmann, P. S. Chen, and E. Kubli. 1991. Ectopic expression of sex peptide alters reproductive behavior of female *D. melanogaster*. Neuron 7:557-563.

Akashi, H. 1999. Inferring the fitness effects of DNA mutations from polymorphism and divergence data: statistical power to detect directional selection under stationarity and free recombination. Genetics 151:221-238.

Andolfatto, P. 2005. Adaptive evolution of non-coding DNA in Drosophila. Nature 437:1149-1152.

Andres, J. A., L. S. Maroja, S. M. Bogdanowicz, W. J. Swanson, and R. G. Harrison. 2006. Molecular evolution of seminal proteins in field crickets. Mol Biol Evol 23:1574-1584.

Anisimova, M., J. P. Bielawski, and Z. Yang. 2001. Accuracy and power of the likelihood ratio test in detecting adaptive molecular evolution. Mol Biol Evol 18:1585-1592.

Asgari, S., G. Zhang, R. Zareie, and O. Schmidt. 2003. A serine proteinase homolog venom protein from an endoparasitoid wasp inhibits melanization of the host hemolymph. Insect Biochem Mol Biol 33:1017-1024.

Begun, D. J., P. Whitley, B. L. Todd, H. M. Waldrip-Dail, and A. G. Clark. 2000. Molecular population genetics of male accessory gland proteins in Drosophila. Genetics 156:1879-1888.

Bertram, M. J., D. M. Neubaum, and M. F. Wolfner. 1996. Localization of the Drosophila male accessory gland protein Acp36DE in the mated female suggests a role in sperm storage. Insect Biochem Mol Biol 26:971-980.

Bloch Qazi, M. C., and M. F. Wolfner. 2003. An early role for the *Drosophila melanogaster* male seminal protein Acp36DE in female sperm storage. J Exp Biol 206:3521-3528.

Carpentier, M., C. Guillemette, J. L. Bailey, G. Boileau, L. Jeannotte, L. DesGroseillers, and J. Charron. 2004. Reduced fertility in male mice deficient in the zinc metallopeptidase NL1. Mol Cell Biol 24:4428-4437.

Carvalho, G. B., P. Kapahi, J. Anderson, and S. Benzer. 2006. Allocrine modulation of feeding behavior by the Sex Peptide of Drosophila. Curr Biol 16: 692-696.

Chapman, T., J. Bangham, G. Vinti, B. Seifried, O. Lung, M. F. Wolfner, H. K. Smith, and L. Partridge. 2003. The sex peptide of *Drosophila melanogaster*: female post-mating responses analyzed by using RNA interference. Proc Natl Acad Sci U S A 100:9923-9928.

Chapman, T., and S. J. Davies. 2004. Functions and analysis of the seminal fluid proteins of male *Drosophila melanogaster* fruit flies. Peptides 25:1477-1490.

Chapman, T., L. F. Liddle, J. M. Kalb, M. F. Wolfner, and L. Partridge. 1995. Cost of mating in *Drosophila melanogaster* females is mediated by male accessory gland products. Nature 373:241-244.

Chintapalli, V. R., J. Wang, and J. A. Dow. 2007. Using FlyAtlas to identify better *Drosophila melanogaster* models of human disease. Nat Genet 39:715-720.

Clark, A. G., M. Aguadé, T. Prout, L. G. Harshman, and C. H. Langley. 1995. Variation in sperm displacement and its association with accessory gland protein loci in *Drosophila melanogaster*. Genetics 139:189-201.

Clark, N. L., J. E. Aagaard, and W. J. Swanson. 2006. Evolution of reproductive proteins from animals and plants. Reproduction 131:11-22.

Clark, N. L., and W. J. Swanson. 2005. Pervasive adaptive evolution in primate seminal proteins. PLoS Genet 1:e35.

Darwin, C. 1871. The Descent of Man, and Selection in Relation to Sex. John Murray, London.

Eberhard, W. G. 1996. Female control: Sexual selection by cryptic female choice. Princeton University Press, Princeton, N. J.

Fay, J. C., and C. I. Wu. 2000. Hitchhiking under positive Darwinian selection. Genetics 155:1405-1413.

Fiumera, A. C., B. L. Dumont, and A. G. Clark. 2005. Sperm competitive ability in *Drosophila melanogaster* associated with variation in male reproductive proteins. Genetics 169:243-257.

Fiumera, A. C., B. L. Dumont, and A. G. Clark. 2007. Associations between sperm competition and natural variation in male reproductive genes on the third chromosome of *Drosophila melanogaster*. Genetics 176:1245-1260.

Gillott, C. 2003. Male accessory gland secretions: Modulators of female reproductive physiology and behavior. Annu Rev Entomol 48: 163-184.

Gupta, S., Y. Wang, and H. Jiang. 2005. *Manduca sexta* prophenoloxidase (proPO) activation requires proPO-activating proteinase (PAP) and serine proteinase homologs (SPHs) simultaneously. Insect Biochem Mol Biol 35:241-248.

Heifetz, Y., O. Lung, E. A. Frongillo, Jr., and M. F. Wolfner. 2000. The Drosophila seminal fluid protein Acp26Aa stimulates release of oocytes by the ovary. Curr Biol 10:99-102.

Herndon, L. A., and M. F. Wolfner. 1995. A Drosophila seminal fluid protein, Acp26Aa, stimulates egg laying in females for 1 day after mating. Proc Natl Acad Sci U S A 92:10114-10118.

Hey, J., and R. M. Kliman. 2002. Interactions between natural selection, recombination and gene density in the genes of Drosophila. Genetics 160:595-608.

Holloway, A. K., and D. J. Begun. 2004. Molecular evolution and population genetics of duplicated accessory gland protein genes in Drosophila. Mol Biol Evol 21:1625-1628.

Huang, T. S., H. Wang, S. Y. Lee, M. W. Johansson, K. Soderhall, and L. Cerenius. 2000. A cell adhesion protein from the crayfish *Pacifastacus leniusculus*, a serine proteinase homologue similar to Drosophila masquerade. J Biol Chem 275:9996-10001.

Hudson, R. R., M. Kreitman, and M. Aguadé. 1987. A test of neutral molecular evolution based on nucleotide data. Genetics 116:153-159.

Hughes, A. L., and M. Nei. 1988. Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. Nature 335:167-170.

Jensen, J. D., Y. Kim, V. B. DuMont, C. F. Aquadro, and C. D. Bustamante. 2005. Distinguishing between selective sweeps and demography using DNA polymorphism data. Genetics 170:1401-1410.

Jiang, H., Y. Wang, X. Q. Yu, and M. R. Kanost. 2003a. Prophenoloxidase-activating proteinase-2 from hemolymph of *Manduca sexta*. A bacteria-inducible serine proteinase containing two clip domains. J Biol Chem 278:3552-3561.

Jiang, H., Y. Wang, X. Q. Yu, Y. Zhu, and M. Kanost. 2003b. Prophenoloxidase-activating proteinase-3 (PAP-3) from *Manduca sexta* hemolymph: a clip-domain serine proteinase regulated by serpin-1J and serine proteinase homologs. Insect Biochem Mol Biol 33:1049-1060.

Kalb, J. M., A. J. DiBenedetto, and M. F. Wolfner. 1993. Probing the function of *Drosophila melanogaster* accessory glands by directed cell ablation. Proc Natl Acad Sci U S A 90:8093-8097.

Kelleher, E. S., W. J. Swanson, and T. A. Markow. 2007. Gene duplication and adaptive evolution of digestive proteases in *Drosophila arizonae* female reproductive tracts. PLoS Genet 3:1541-1549.

Kern, A. D., C. D. Jones, and D. J. Begun. 2004. Molecular population genetics of male accessory gland proteins in the *Drosophila simulans* complex. Genetics 167:725-735.

Kreitman, M. 1983. Nucleotide polymorphism at the alcohol dehydrogenase locus of *Drosophila melanogaster*. Nature 304:412-417.

Kubli, E. 2003. Sex-peptides: seminal peptides of the Drosophila male. Cell Mol Life Sci 60:1689-1704.

Kwon, T. H., M. S. Kim, H. W. Choi, C. H. Joo, M. Y. Cho, and B. L. Lee. 2000. A masquerade-like serine proteinase homologue is necessary for phenoloxidase activity in the coleopteran insect, *Holotrichia diomphalia* larvae. Eur J Biochem 267:6188-6196.

Lawniczak, M. K., and D. J. Begun. 2007. Molecular Population Genetics of Female-expressed Mating-induced Serine Proteases in *Drosophila melanogaster*. Mol Biol Evol.

Lee, K. Y., R. Zhang, M. S. Kim, J. W. Park, H. Y. Park, S. Kawabata, and B. L. Lee. 2002. A zymogen form of masquerade-like serine proteinase homologue is cleaved during pro-phenoloxidase activation by Ca2+ in coleopteran and Tenebrio molitor larvae. Eur J Biochem 269:4375-4383.

Ligoxygakis, P., N. Pelte, J. A. Hoffmann, and J. M. Reichhart. 2002a. Activation of Drosophila Toll during fungal infection by a blood serine protease. Science 297:114-116.

Ligoxygakis, P., N. Pelte, C. Ji, V. Leclerc, B. Duvic, M. Belvin, H. Jiang, J. A. Hoffmann, and J. M. Reichhart. 2002b. A serpin mutant links Toll activation to melanization in the host defence of Drosophila. EMBO J 21:6330-6337.

Lin, C. Y., K. Y. Hu, S. H. Ho, and Y. L. Song. 2006. Cloning and characterization of a shrimp clip domain serine protease homolog (c-SPH) as a cell adhesion molecule. Dev Comp Immunol 30:1132-1144.

Liu, H., and E. Kubli. 2003. Sex-peptide is the molecular basis of the sperm effect in *Drosophila melanogaster*. Proc Natl Acad Sci U S A 100:9929-9933.

Lung, O., L. Kuo, and M. F. Wolfner. 2001. Drosophila males transfer antibacterial proteins from their accessory gland and ejaculatory duct to their mates. J Insect Physiol 47:617-622.

Lung, O., U. Tram, C. M. Finnerty, M. A. Eipper-Mains, J. M. Kalb, and M. F. Wolfner. 2002. The *Drosophila melanogaster* seminal fluid protein Acp62F is a protease inhibitor that is toxic upon ectopic expression. Genetics 160:211-224.

Mack, P. D., A. Kapelnikov, Y. Heifetz, and M. Bender. 2006. Mating-responsive genes in reproductive tissues of female *Drosophila melanogaster*. Proc Natl Acad Sci U S A 103:10358-10363.

Mbikay, M., H. Tadros, N. Ishida, C. P. Lerner, E. De Lamirande, A. Chen, M. El-Alfy, Y. Clermont, N. G. Seidah, M. Chretien, C. Gagnon, and E. M. Simpson. 1997. Impaired fertility in mice deficient for the testicular germ-cell protease PC4. Proc Natl Acad Sci U S A 94:6842-6846.

McConnell, T. J., W. S. Talbot, R. A. McIndoe, and E. K. Wakeland. 1988. The origin of MHC class II gene polymorphism within the genus Mus. Nature 332:651-654.

McDonald, J. H., and M. Kreitman. 1991. Adaptive protein evolution at the Adh locus in Drosophila. Nature 351:652-654.

Monsma, S. A., and M. F. Wolfner. 1988. Structure and expression of a Drosophila male accessory gland gene whose product resembles a peptide pheromone precursor. Genes Dev 2:1063-1073.

Mueller, J. L., J. L. Page, and M. F. Wolfner. 2007. An ectopic expression screen reveals the protective and toxic effects of Drosophila seminal fluid proteins. Genetics 175:777-783.

Mueller, J. L., K. R. Ram, L. A. McGraw, M. C. Bloch Qazi, E. D. Siggia, A. G. Clark, C. F. Aquadro, and M. F. Wolfner. 2005. Cross-species comparison of Drosophila male accessory gland protein genes. Genetics 171:131-143.

Mueller, J. L., D. R. Ripoll, C. F. Aquadro, and M. F. Wolfner. 2004. Comparative structural modeling and inference of conserved protein classes in Drosophila seminal fluid. Proc Natl Acad Sci U S A 101:13542-13547.

Murer, V., J. F. Spetz, U. Hengst, L. M. Altrogge, A. de Agostini, and D. Monard. 2001. Male fertility defects in mice lacking the serine protease inhibitor protease nexin-1. Proc Natl Acad Sci U S A 98:3029-3033.

Neubaum, D. M., and M. F. Wolfner. 1999. Mated *Drosophila melanogaster* females require a seminal fluid protein, Acp36DE, to store sperm efficiently. Genetics 153:845-857.

Nie, G., Y. Li, M. Wang, Y. X. Liu, J. K. Findlay, and L. A. Salamonsen. 2005. Inhibiting uterine PC6 blocks embryo implantation: an obligatory role for a proprotein convertase in fertility. Biol Reprod 72:1029-1036.

Panhuis, T. M., N. L. Clark, and W. J. Swanson. 2006. Rapid evolution of reproductive proteins in abalone and Drosophila. Philos Trans R Soc Lond B Biol Sci 361:261-268.

Panhuis, T. M., and W. J. Swanson. 2006. Molecular evolution and population genetic analysis of candidate female reproductive genes in Drosophila. Genetics 173:2039-2047.

Park, M., and M. F. Wolfner. 1995. Male and female cooperate in the prohormone-like processing of a *Drosophila melanogaster* seminal fluid protein. Dev Biol 171:694-702.

Poiani, A. 2006. Complexity of seminal fluid: A review. Behav Evol Sociobiol 60: 289-310.

Pond, S. K., and S. V. Muse. 2005. Site-to-site variation of synonymous substitution rates. Mol Biol Evol 22:2375-2385.

Pond, S. L., S. D. Frost, and S. V. Muse. 2005. HyPhy: hypothesis testing using phylogenies. Bioinformatics 21:676-679.

Pool, J. E., and C. F. Aquadro. 2006. History and structure of sub-Saharan populations of *Drosophila melanogaster*. Genetics 174:915-929.

Ravi Ram, K., L. K. Sirot, and M. F. Wolfner. 2006. Predicted seminal astacin-like protease is required for processing of reproductive proteins in *Drosophila melanogaster*. Proc Natl Acad Sci U S A 103:18674-18679.

Ravi Ram, K., and M. F. Wolfner. 2007. Seminal influences: *Drosophila* Acps and the molecular interplay between males and females during reproduction. Integrative and Comparative Biology 47: 427-445.

Ross, J., H. Jiang, M. R. Kanost, and Y. Wang. 2003. Serine proteases and their homologs in the *Drosophila melanogaster* genome: an initial analysis of sequence conservation and phylogenetic relationships. Gene 304:117-131.

Rozas, J., J. C. Sanchez-DelBarrio, X. Messeguer, and R. Rozas. 2003. DnaSP, DNA polymorphism analyses by the coalescent and other methods. Bioinformatics 19:2496-2497.

Samakovlis, C., P. Kylsten, D. A. Kimbrell, A. Engstrom, and D. Hultmark. 1991. The andropin gene and its product, a male-specific antibacterial peptide in *Drosophila melanogaster*. Embo J 10:163-169.

Saudan, P., K. Hauck, M. Soller, Y. Choffat, M. Ottiger, M. Sporri, Z. Ding, D. Hess, P. M. Gehrig, S. Klauser, P. Hunziker, and E. Kubli. 2002. Ductus ejaculatorius peptide 99B (DUP99B), a novel *Drosophila melanogaster* sex-peptide pheromone. Eur J Biochem 269:989-997.

Schully, S. D., and M. E. Hellberg. 2006. Positive Selection on Nucleotide Substitutions and Indels in Accessory Gland Proteins of the *Drosophila pseudoobscura* Subgroup. J Mol Evol.

Sim, R. B., and S. A. Tsiftsoglou. 2004. Proteases of the complement system. Biochem Soc Trans 32:21-27.

Simmons, L. 2001. Sperm competition and its evolutionary consequences. Princeton University Press, Princeton.

Stevison, L. S., B. A. Counterman, and M. A. Noor. 2004. Molecular evolution of X-linked accessory gland proteins in *Drosophila pseudoobscura*. J Hered 95:114-118.

Swanson, W. J., A. G. Clark, H. M. Waldrip-Dail, M. F. Wolfner, and C. F. Aquadro. 2001. Evolutionary EST analysis identifies rapidly evolving male reproductive proteins in Drosophila. Proc Natl Acad Sci U S A 98:7375-7379.

Swanson, W. J., and V. D. Vacquier. 2002. The rapid evolution of reproductive proteins. Nat Rev Genet 3:137-144.

Swanson, W. J., A. Wong, M. F. Wolfner, and C. F. Aquadro. 2004. Evolutionary expressed sequence tag analysis of Drosophila female reproductive tracts identifies genes subjected to positive selection. Genetics 168:1457-1465.

Tajima, F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 123:585-595.

Thornton, K. R., J. D. Jensen, C. Becquet, and P. Andolfatto. 2007. Progress and prospects in mapping recent selection in the genome. Heredity 98:340-348.

Tram, U., and M. F. Wolfner. 1998. Seminal fluid regulation of female sexual attractiveness in *Drosophila melanogaster*. Proc Natl Acad Sci U S A 95:4051-4054.

Tsaur, S. C., C. T. Ting, and C. I. Wu. 1998. Positive selection driving the evolution of a gene of male reproduction, Acp26Aa, of Drosophila: II. Divergence versus polymorphism. Mol Biol Evol 15:1040-1046.

Tsaur, S. C., and C. I. Wu. 1997. Positive selection and the molecular evolution of a gene of male reproduction, Acp26Aa of Drosophila. Mol Biol Evol 14:544-549.

Uhrin, P., C. Schofer, J. Zaujec, L. Ryban, M. Hilpert, K. Weipoltshammer, I. Jerabek, I. Pirtzkall, M. Furtmuller, M. Dewerchin, B. R. Binder, and M. Geiger. 2007.

Male fertility and protein C inhibitor/plasminogen activator inhibitor-3 (PCI): localization of PCI in mouse testis and failure of single plasminogen activator knockout to restore spermatogenesis in PCI-deficient mice. Fertil Steril.

Wagstaff, B. J., and D. J. Begun. 2005a. Comparative genomics of accessory gland protein genes in *Drosophila melanogaster* and *D. pseudoobscura*. Mol Biol Evol 22:818-832.

Wagstaff, B. J., and D. J. Begun. 2005b. Molecular Population Genetics of Accessory Gland Protein Genes and Testis-expressed Genes in Drosophila mojavensis and D. arizonae. Genetics 171: 1083-1101.

Watorek, W. 2003. Azurocidin -- inactive serine proteinase homolog acting as a multifunctional inflammatory mediator. Acta Biochim Pol 50:743-752.

Wigby, S., and T. Chapman. 2005. Sex peptide causes mating costs in female *Drosophila melanogaster*. Curr Biol 15:316-321.

Wolfner, M. F. 2002. The gifts that keep on giving: physiological functions and evolutionary dynamics of male seminal proteins in Drosophila. Heredity 88:85-93.

Wolfner, M. F., Y. Heifetz, and S. W. Applebaum. 2005. Gonadal glands and their gene products *in* L. I. Gilbert, K. Iatrou, and S. S. Gill, eds. Comprehensive molecular insect science: Reproduction and Development. Elsevier Ltd., Oxford.

Wong, A., S. N. Albright, and M. F. Wolfner. 2006. Evidence for structural constraint on ovulin, a rapidly evolving *Drosophila melanogaster* seminal protein. Proc Natl Acad Sci U S A 103:18644-18649.

Wong, A., A. D. Jensen, J. E. Pool, and C. F. Aquadro. 2007. Phylogenetic incongruence in the *Drosophila melanogaster* species group. Mol Phy Evol 43: 1138-1150.

Wright, S. I., and B. Charlesworth. 2004. The HKA test revisited: a maximum-likelihood-ratio test of the standard neutral model. Genetics 168:1071-1076.

Xue, L., and M. Noll. 2000. Drosophila female sexual behavior induced by sterile males showing copulation complementation. Proc Natl Acad Sci U S A 97:3272-3275.

Yang, Z. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. Comput Appl Biosci 13:555-556.

Yang, Z., R. Nielsen, N. Goldman, and A. M. Pedersen. 2000. Codon-substitution models for heterogeneous selection pressure at amino acid sites. Genetics 155:431-449.

Yu, X. Q., H. Jiang, Y. Wang, and M. R. Kanost. 2003. Nonproteolytic serine proteinase homologs are involved in prophenoloxidase activation in the tobacco hornworm, *Manduca sexta*. Insect Biochem Mol Biol 33:197-208.

CHAPTER 5

EVIDENCE FOR MOLECULAR CO-EVOLUTION BETWEEN THE SEXES IN *DROSOPHILA MELANOGASTER*

Introduction

Interactions between males and females are thought to underlie the rapid evolution of many morphological and behavioral traits involved in mating and reproduction. Male –male competition and female mating preferences are powerful forces in driving the evolution of male display traits (Darwin 1871; Andersson 1994), and the morphology and chemical environment of the female reproductive tract can exert selection even after mating has ended (e.g., Eberhard 1996; Miller and Pitnick 2002; Pitnick et al. 2003).

Pre- and post-mating sexual selection are also thought to operate on molecules involved in mating and fertilization. Prior to mating, for example, males and/or females of many animal species use pheromonal cues in the process of mate selection, with individuals within a species showing preferences for different pheromones or pheromone blends. In the Australian fruitfly *Drosophila serrata*, for example, male and female contact-pheromone blends change rapidly in response to sexual selection (e.g., Chenoweth, Rundle, and Blows 2008; Higgie and Blows 2008). Post-mating, various components of the male ejaculate may aid in competition between sperm from different males, and females may discriminate between different males on the same basis (Keller and Reeve 1995; Eberhard 1996; Ball and Parker 2003; Cameron, Day, and Rowe 2007). In *Drosophila melanogaster*, for example, genotype at several loci encoding accessory gland proteins (Acps), which form a major component of the male ejaculate, predicts a male's success in sperm competition assays (Clark et al. 1995; Fiumera, Dumont, and Clark 2005; Fiumera, Dumont, and Clark 2007). The success of

a particular male in sperm competition is dependent on his own genotype (Clark et al. 1995; Civetta and Clark 2000; Fiumera, Dumont, and Clark 2005; Fiumera, Dumont, and Clark 2007) and that of his mate (Clark, Begun, and Prout 1999; Civetta and Clark 2000), suggesting that both males and females contribute to the outcome of sperm competition.

The Acps of *D. melanogaster* constitute a powerful system for studying post-copulatory sexual selection. Acps are produced in the male accessory glands, and are transferred to the female during copulation, along with sperm and other secretions. In addition to their roles in sperm competition, Acps are required for the induction of a number of female post-mating responses, including increased rates of egg-laying and feeding, sperm storage, a reduced propensity to remate, and changes in uterine morphology (Adams and Wolfner 2007; Ravi Ram and Wolfner 2007). In addition, Acps contribute to the 'cost-of-mating', whereby mated Drosophila females suffer reduced lifespans in comparison to virgin females (Chapman et al. 1995; Wigby and Chapman 2005).

Two major lines of evidence suggest that Acps are subject to post-copulatory sexual selection. First, like many morphological traits involved in mating and reproduction, a number of Acps evolve rapidly; population genetic and molecular evolutionary analyses indicate that this rapid evolution is driven in several cases by positive selection, rather than relaxed constraint. Positive selection has been documented on Acps involved in ovulation (Ovulin; Aguadé, Miyashita, and Langley 1992; Tsaur and Wu 1997; Aguadé 1998; Tsaur, Ting, and Wu 1998) and sperm storage (Acp36DE – Begun et al. 2000; CG9997 – Wong et al. 2008, Acp29AB – Aguadé 1999), and Acps as a class are more likely to show evidence for positive selection than are non-reproductive proteins (Haerty et al. 2007). Second, comparisons between rates of Acp evolution in taxa with different mating habits have found that

selection is stronger in species that mate more frequently (Wagstaff and Begun 2005b; Wagstaff and Begun 2007). This result suggests that sperm competition and/or female sperm preference following multiple mating plays an important role in Acp evolution.

Less is known concerning the relative roles played by male-male and male-female interactions in Acp evolution. If male-female co-evolution drives the evolution of some Acps, as has been widely proposed (e.g., Aguadé, Miyashita, and Langley 1992; Civetta and Singh 1995; Tsaur and Wu 1997; Swanson and Vacquier 2002; Lawniczak and Begun 2007; Wong et al. 2008), then a simple prediction is that female interactors of Acps should also evolve rapidly. Moreover, such interactors may show other evidence of co-evolution, such as LD with interacting Acps and correlated patterns of substitution (Kirkpatrick 1982; Dimmic et al. 2005). Only one female receptor for an Acp has been identified, SPR, the receptor for sex-peptide (Yapici et al. 2008). There is some evidence for non-neutral evolution at the sex-peptide locus, although this may result from demographic factors rather than selection (Cirera and Aguadé 1997); co-evolutionary studies may be of value, especially if further work finds convincing evidence for positive selection on sex-peptide. A host of other potential female partners for Acps have, however, been identified. Several studies have identified genes expressed in the somatic portion of the female reproductive tract. Many of these genes are predicted to encode extracellular or transmembrane proteins, which could potentially interact with Acps after mating (Swanson et al. 2004; Mack et al. 2006; Kelleher, Swanson, and Markow 2007; Allen and Spradling 2008). In addition, several microarray studies have identified genes upregulated after mating, in some cases in the female reproductive tract, although proteins coded by this group of genes may not interact directly with Acps (Lawniczak and Begun 2004; McGraw et al. 2004; Mack et al. 2006). A handful of genes identified in these studies (primarily protease-encoding) show evidence for positive selection (Swanson et al. 2004; Panhuis

and Swanson 2006; Lawniczak and Begun 2007), raising the possibility of male-female co-evolution. Nonetheless, a large scale comparison found that 679 female reproductive tract genes identified by Swanson et al. 2004 and Mack et al. 2006 are no more likely to undergo positive selection, and in fact evolve more slowly at the amino acid level, in comparison to putative non-reproductive genes (Haerty et al. 2007).

In addition to post-copulatory sexual selection, immune pressures may contribute to the evolution of both male and female reproductive tract proteins. Host-pathogen interactions are thought to be a common cause of rapid protein evolution (e.g., Sackton et al. 2007). Sexually transmitted diseases have not to my knowledge been documented in Drosophila, but pathogen introduction during mating is a plausible risk. Several Acps appear to have anti-bacterial activity (Lung, Kuo, and Wolfner 2001; Mueller, Page, and Wolfner 2007), and genes with known roles in immunity are expressed in the reproductive tracts of both males and females (Table 5.1). Mating alters the expression levels of several anti-microbial peptides in females (Lawniczak and Begun 2004; McGraw et al. 2004; Peng, Zipperlen, and Kubli 2005; Domanitskaya et al. 2007), although the physiological consequences of these gene expression changes are not clear (Wigby et al. 2008). Together, these observations raise the possibility that host-pathogen interactions in the female reproductive tract could also contribute to rapid Acp evolution (see also Lawniczak et al. 2007).

In this study, we address the causes of rapid reproductive protein evolution in *Drosophila* using a set of genes encoding potentially interacting reproductive molecules: Proteolysis regulators (proteases and modulators of their activity) and targets of proteolysis present in the male accessory gland or in the female reproductive tract of *D. melanogaster*. Proteolysis is thought to play an important role in regulating the activity of Acps. Proteolytic cleavage of sex-peptide (SP), for example, releases its bioactive C-terminal peptide from sperm, allowing this peptide to enter the

hemolymph and mediate several long-term post-mating responses (Peng et al. 2005). In addition, several putative cleavage products of ovulin are capable of inducing ovulation, suggesting that proteolysis of ovulin releases active peptide hormones (Heifetz et al. 2005). At least one protease produced in the accessory gland is necessary for ovulin cleavage (Ravi Ram, Sirot, and Wolfner 2006), and it is thought that female factors are also required (Park and Wolfner 1995).

Previous studies have identified an abundance of proteolysis regulators expressed in the male accessory gland (Swanson et al. 2001; Mueller et al. 2004; FlyAtlas.org) and in the somatic portions of the female reproductive tract (Swanson et al. 2004; Mack et al. 2006; Kelleher, Swanson, and Markow 2007; Lawniczak and Begun 2007; Allen and Spradling 2008). Since proteases, modulators of protease activity, and targets of proteolysis could potentially form interacting networks of proteins, and are suitably localized for interaction following mating, we propose that these molecules represent good candidate interactors for studying co-evolution. In this study, we use polymorphism data from 37 proteolysis regulators and 3 targets of proteolysis to test key predictions of molecular co-evolution. We find evidence in support of co-evolution, with similar levels of positive selection on male- and female-reproductive tract genes, and elevated linkage disequilibrium between genes expressed in different sexes. In addition, several genes subject to positive selection have documented roles in immunity, suggesting an important role for host-pathogen interactions in reproductive protein evolution.

Materials and Methods

Loci

We surveyed polymorphism at 3 loci encoding known targets of proteolysis, and 37 loci encoding proteolysis regulators - predicted proteases, protease inhibitors

(PIs), or protease homologs. Protease homologs resemble proteases in primary sequence and tertiary structure, but carry one or more catalytic site mutations such that they probably lack normal catalytic activity. Nonetheless, protease homologs have been reported to modulate protease activity, either as agonists or antagonists (e.g., Kwon et al. 2000; Lee et al. 2002; Asgari et al. 2003; Gupta, Wang, and Jiang 2005). Of these 40 loci, 13 (6 PIs and 7 protease/protease homologs) are known to be expressed in the somatic portion of the female reproductive tract, and 27 (7 PIs, 17 protease/protease homologs, and 3 targets) have strongly male accessory gland biased expression. It should be noted that the degree of tissue specificity differs substantially between the male and female samples: the male accessory gland genes were selected on the basis of strong expression bias (Swanson et al. 2001; FlyAtlas.org), and microarray studies examining 11 adult tissues support their high specificity (FlyAtlas.org). The female reproductive tract genes, by contrast, have varying degrees of tissue specificity, and it is not known whether any are specifically expressed in the female reproductive tract.

Previous studies have identified six Acps that undergo proteolysis following transfer to the female: sex-peptide, ovulin, the sperm storage protein Acp36DE, the protease CG11864, the protease homolog CG9997, and the protease inhibitor CG9334. Here, we use the term "targets" to refer only to the first three of the six known targets of proteolysis, with the latter three considered under proteases or protease inhibitors, respectively. Table 5.1 lists all 40 loci, with predicted molecular functions and known biological roles.

Drosophila strains and DNA sequencing

For polymorphism based analyses in *D. melanogaster*, we used chromosome extraction lines for the X, 2nd, and 3rd chromosomes, isolated from isofemale lines

Table 5.1 Genes surveyed in chapter 5. Targets are proteins known to undergo proteolysis following mating. PI: Predicted protease inhibitors. Prot.: Predicted catalytic proteases. Prot. hom.: Predicted protease homologs.

Gene	Type	Function/Effects	Sex
CG8982 (Acp26Aa, ovulin)	Target	Ovulation	M
CG7157 (Acp36DE)	Target	Sperm storage	M
CG17673 (Acp70A, sex-peptide)	Target	Remating, egg-production and laying, feeding	M
CG1262 (Acp62F)	PI	Sperm competition, toxic	M
CG1342	PI	Unknown	M
CG8137	PI	Toxic	M
CG9334	PI	Immunity	M
CG10956	PI	Unknown	M
CG31902	PI	Unknown	M
CG32203	PI	Unknown	M
CG33121	PI	Unknown	M
CG9997	Prot. hom.	Remating, egg-production and laying, sperm release	M
CG11864	Prot.	Cleavage of ovulin	M
CG6168	Prot.	Immunity	M
CG32382	Prot. hom.	Immunity	M
CG32383	Prot. hom.	Immunity	M
CG1895	Prot.	Unknown	M
CG6069	Prot. hom.	Unknown	M
CG10586	Prot.	Unknown	M
CG10587	Prot.	Unknown	M
CG11664	Prot. hom.	Unknown	M
C13518	Prot.	Unknown	M
CG17242	Prot.	Unknown	M
CG18557	Prot.	Unknown	M
CG31681	Prot.	Unknown	M
CG32833	Prot.	Unknown	M
CG1857 (*necrotic*)	PI	Immunity	F
CG11331	PI	Immunity	F
CG1865	PI	Unknown	F
CG3604	PI	Unknown	F
CG9456	PI	Unknown	F + M
CG18525	PI	Unknown	F
CG3066	Prot.	Immunity	F
CG3074	Prot.	Eggshell matrix	F
CG3097	Prot.	Unknown	F
CG9849	Prot.	Unknown	F
CG13318	Prot. hom.	Unknown	F
CG18125	Prot.	Unknown	F
CG31199	Prot.	Unknown	F

derived from a Ugandan population (population samples are described in Pool and Aquadro 2006). *D. simulans* sequences were collected from isofemale lines derived from a Madagascar population.

DNA was extracted using the Puregene DNA purification kit (Gentra Systems, Minneapolis, MN). Genes were amplified by polymerase chain reaction (PCR), and PCR products were sequenced using BigDye chemistry (Applied Biosystems, Foster City, CA) on an ABI 3730 automated sequencer at the Cornell University Life Sciences Core Laboratories Center. PCR and sequencing primer sequences are available upon request. Sequence alignments were performed using the ClustalW algorithm as implemented in CodonCode Aligner (CodonCode Corp., Dedham, MA).

Molecular population genetics

Summary statistics (π, Tajima's D) were calculated using the Analysis software package, which is based on the libsequence C++ libraries (Thornton 2003). For inferences of selection at loci encoding putative proteolysis regulators and targets of proteolysis, we used the mkprf method of (Bustamante et al. 2002; Barrier et al. 2003). mkprf is a Bayesian method that increases the power of the traditional McDonald-Kreitman test (McDonald and Kreitman 1991) by using information from multiple loci to infer the time since divergence of two species (τ), a parameter shared between all loci. mkprf assumes that the distribution of selection coefficients ($\gamma = 2N_eS$) on amino acid changes for a given class of genes follows a Gaussian distribution, such that the distributions of selection coefficients for two or more classes of genes can be compared using the posterior distributions of γ. In addition, mkprf provides an estimate of γ at each locus; a γ significantly greater than 0 at a given locus constitutes evidence for positive selection on amino acid changes between species (assuming neutrality of synonymous changes). Here, we present mkprf analyses using

polymorphism data from *D. melanogaster* alone, with a single *D. simulans* sequence used for divergence; the current release of mkprf allows polymorphism data from only a single species. mkprf assumes demographic stationarity; this assumption is more likely to be true of the *D. melanogaster* population sampled here than of the *D. simulans* population (see below). The mkprf analysis was run on the Cornell University Computational Biology Service Unit cluster (partially funded by Microsoft Corporation). Ten chains were run for 10000 iterations, with the first 1000 iterations discarded for burn-in. Default settings were used for all other parameters.

Inter-locus linkage disequilibrium parameters were estimated using a custom Perl script (available upon request). For every pair of loci, the statistics D, D', and r^2 were calculated for each pair of polymorphic sites (each with the minor allele represented at least twice in the sample). For each pair of loci, ZnS, the average of all pairwise values of r^2, was used as a summary measure of inter-locus LD (Kelly 1997). Pairs of genes with a low number of site pairs were excluded from the analysis; the bottom quartile was excluded for this reason in each species, such that only gene pairs with >2 site pairs were considered in *D. melanogaster*, and only gene pairs with >8 site pairs were considered in *D. simulans.* We furthermore excluded pairs of genes sequenced in less than 6 strains in common. Polymorphisms were unpolarized with respect to being ancestral or derived; further analyses using polarized changes are underway.

Results

Patterns of diversity

We sequenced 40 loci encoding known targets of proteolysis and putative proteolysis regulators in population samples of *D. melanogaster* and *D. simulans*. An average of 14.1 and 12.6 alleles were sequenced for each locus in the *D. melanogaster*

and *D. simulans* samples, respectively. Over all loci, average diversity in *D. melanogaster* (π) was 0.0071 (sd = 0.0046). Diversity was substantially higher in *D. simulans*, where mean $\pi = 0.015$ (sd = 0.0061). Both estimates are similar to those previously documented in the literature (e.g., Andolfatto 2005; Begun et al. 2007). The difference in diversity between the two species was highly significant (Figure 5.1A; paired T-test $P = 3.3 \times 10^{-8}$), consistent with a larger effective population size in *D. simulans*, in contrast to inferences from a recent study (Nolte and Schlotterer 2008). Systematic differences were also observed in the site frequency spectrum between the two species, with a significantly lower Tajima's D in *D. simulans* indicating a relative excess of rare alleles in that species (Figure 5.1B; paired T-test = 0.00013).

Inferences of selection

Previous multi-locus studies have suggested that genes expressed in the female reproductive tract of *D. melanogaster* evolve more slowly on average between species, and undergo less positive selection, than do male reproductive tract genes (Haerty et al. 2007). We used mkprf to compare selective pressures on putative proteolysis regulators and targets of proteolysis, which represent potentially interacting male- and female- reproductive tract proteins. In our analysis, we used information from 39 loci to estimate τ, but allowed separate distributions of γ for male- and female-reproductive tract genes (Figure 5.2A). We were unable to include the fortieth gene, CG9334, in this analysis, because it appears to be a pseudogene in *D. simulans* (see below). For both classes of gene, we estimated that the mean selection coefficient on amino acid changes was greater than 0, suggesting that non-lethal amino acid mutations are beneficial on average (females: mean $\gamma = 2.00$, sd = 0.81; males: mean $\gamma = 1.07$, sd = 0.44). In our dataset, mean γ is higher for the female reproductive tract genes, although overlapping standard deviations suggest that the distributions are not

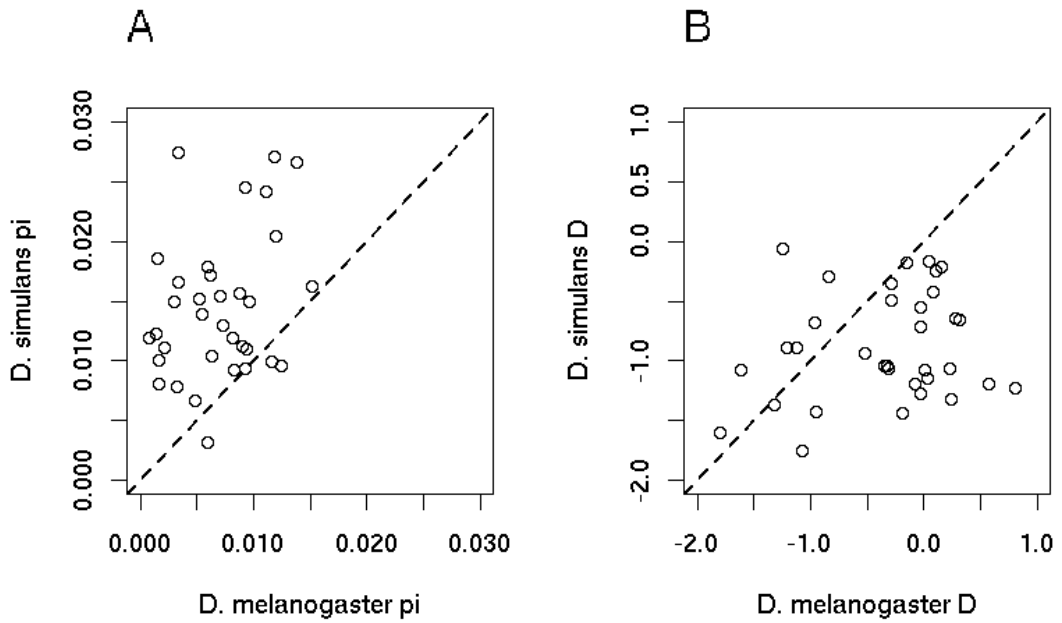
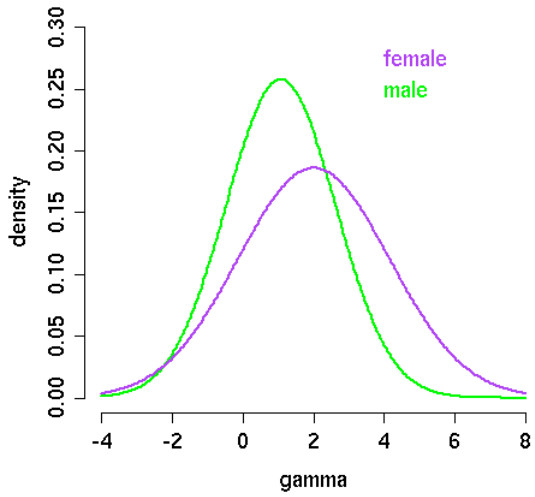


Figure 5.1 Summaries of diversity (π; A) and the site frequency spectrum (Tajima's D; B) for 37 proteolysis regulators and 3 targets of proteolysis. Values of π or D in *D. simulans* are plotted against those in *D. melanogaster*. The dashed line in each plot has a slope of 1, such that genes falling above it have a higher value of π (or D) in *D. simulans*, and genes falling below it have a lower value in *D. simulans*. Note that π tends to be higher, and D lower, in *D. simulans*.

A



B

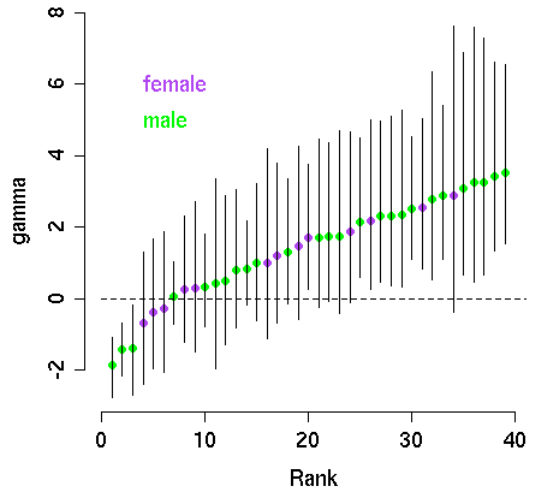


Figure 5.2 Estimates of the selection coefficient γ ($2N_eS$) on non-lethal amino acid changes in *D. melanogaster*. (A) Estimated distributions of γ for proteolysis regulators and targets expressed in the male- (green) and female- (purple) reproductive tracts of *D. melanogaster*. (B) Estimates of γ for each gene, with 95% confidence intervals. The dashed line marks γ = 0 (no selection).

significantly different. Thus, in contrast to previous studies, and consistent with the hypothesis of male-female co-evolution, we find similar levels of selection on candidate male- and female- reproductive genes.

Of the 39 reproductive tract genes that we examined, 19 (14 male and 5 female) show evidence for positive selection (Figure 5.2B; Tables 5.2 and 5.3). A larger fraction of male reproductive tract genes shows evidence for positive selection in comparison to female reproductive tract genes (48% vs. 38%), but this difference is not significant (Fisher's exact test 2-tailed $P = 0.739$). Positive selection has been previously documented on five of these genes (CG7157/Acp36DE - Begun et al. 2000; CG8982/Acp26Aa/Ovulin – Aguadé, Miyashita, and Langley 1992; Tsaur and Wu 1997; Tsaur, Ting, and Wu 1998; CG9997 – Wong et al. 2008; CG3066 – Swanson et al. 2004; CG18125 – Lawniczak and Begun 2007), but the remaining 14 genes are novel in this respect. Interestingly, roles assigned to positively selected genes in genetic and transgenic studies (Tables 5.2 and 5.3) include both reproductive functions (sperm storage, ovulation) and functions in immunity.

Linkage disequilibrium

Models of mate choice predict linkage disequilibrium between trait and preference loci (e.g., Kirkpatrick 1982). By analogy, we expected that loci involved in post-copulatory sexual selection should show elevated linkage disequilibrium (LD). Specifically, if our datasets are in fact enriched for genes encoding interacting male and female reproductive proteins, then we predicted an excess of LD in comparisons involving genes expressed in different sexes (male-female) over comparisons within sexes (male-male or female-female). It should be noted, however, that interactions within sexes could also generate LD, and that some such interactions are known in male ejaculates (Ravi Ram, Sirot, and Wolfner 2006).

Table 5.2 Inferences of selection on male accessory gland genes by mkprf γ is the mean selection coefficient ($2N_eS$) on non-lethal amino acid changes for a given gene. A γ significantly greater than zero (fifth column) is indicative of positive selection.

Gene	Type	Function/effects	Mean γ (S.D.)	*P* (γ<0)
CG7157	target	Sperm storage	2.50 (0.88)	0
CG8982	target	Ovulation	3.51 (1.29)	0
CG32203	PI		2.88 (1.10)	0.00001
CG8137	PI	Toxic	3.43 (1.36)	0.00002
CG33121	PI		2.15 (1.01)	0.00088
CG9997	prot	Sperm release, remating, egg-laying	3.10 (1.61)	0.00234
CG32383	prot	Immunity	3.27 (1.70)	0.00271
CG32382	prot	Immunity	2.79 (1.50)	0.00339
CG17242	prot		2.31 (1.16)	0.00376
CG10586	prot		2.31 (1.21)	0.0064
CG6069	prot		3.26 (1.84)	0.00666
CG11664	prot		2.35 (1.27)	0.00875
CG1262	PI	Sperm competition, toxic	1.75 (1.14)	0.03285
CG10956	PI		1.30 (0.90)	0.04604

Table 5.3 Inferences of selection on female reproductive tract genes by mkprf

Gene	Type	Function/effects	Mean γ (S.D.)	*P* (γ<0)
CG1857 (*nec*)	PI	Immunity	2.54 (1.07)	0.00033
CG18125	prot	Mating induced - spermathecae	1.70 (0.90)	0.0075
CG3066	prot	Immunity	2.18 (1.21)	0.01072
CG3074	prot	Eggshell matrix	1.87 (1.22)	0.0334
CG9849	prot		2.90 (2.06)	0.04659

In our *D. melanogaster* population sample, we find no evidence for an excess of LD in male-female comparisons (Figure 5.3A); in fact, ZnS tends to be slightly higher in male-male comparisons than in male-female or female-female comparisons, although this difference is not significant (t = 1.36; $P = 0.176$). In *D. simulans*, however, we find a significant excess of LD in male-female comparisons over female-female comparisons (Figure 5.3B; t = 2.56, $P = 0.011$), consistent with our prediction. We additionally find a slight excess of LD in male-male comparisons over female-female comparisons (t = 2.05, $P = 0.04$), perhaps suggestive of interactions between male derived molecules.

Gene pairs with particularly high ZnS represent good candidates for interaction studies. In Figure 5.3, points lying outside the "whiskers" on each box-plot represent potential outliers. Interestingly, several outliers involve known targets of proteolysis. In the *D. melanogaster* dataset, for example, one of the outlier datapoints represents the comparison between the male sperm storage protein Acp36DE and the female protease CG3074. Interestingly, both of these genes also showed evidence for positive selection (Tables 5.2 and 5.3). A genetic interaction could be tested for by examining the cleavage of Acp36DE in CG3074 knockdown or knockout females (although other interactions are possible for which no cleavage phenotype would be observed).

The difference between the *D. melanogaster* and *D. simulans* datasets may reflect differences in power and in sampling. Because the *D. melanogaster* data were collected from extracted chromosome lines representing only partially overlapping sets of strains, sample sizes tended to be smaller for comparisons between loci on different chromosomes. Small sample sizes ($n < 6$) were discarded from our analyses, and so many comparisons are absent in the *melanogaster* dataset. Moreover, higher nucleotide diversity in the *D. simulans* population sample provides greater power, since more site pairs are tested for each pair of genes.

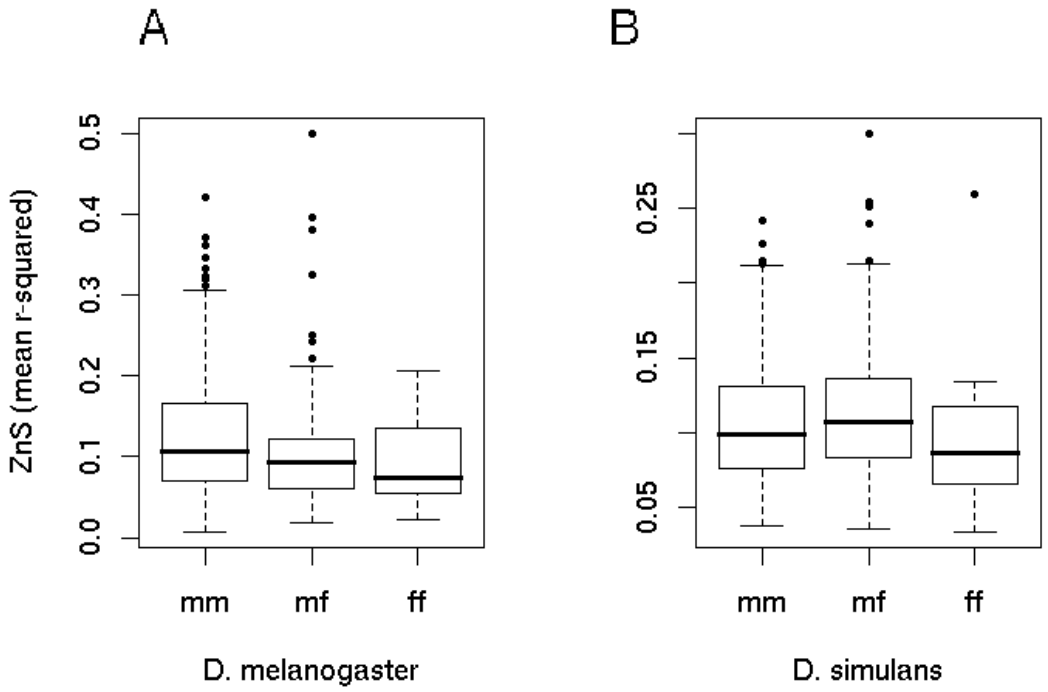


Figure 5.3. Linkage disequilibrium between genes expressed in the male and female reproductive tracts, for *D. melanogaster* (A) and *D. simulans* (B). ZnS is the average of r^2 for all site pairs in a given comparison. "mm": comparisons between Acp loci. "mf": comparisons between genes expressed in the male and female reproductive tracts. "ff": comparisons between female reproductive tract genes. On each boxplot, the central horizontal line indicates the median, the edges of the box represent the quartiles, and points outside of the whiskers represent potential outliers.

Segregating and fixed putative loss-of-function alleles – Gene death in action?

Previous studies have suggested that Acps tend to turn over rapidly between species. Orthologs to *D. melanogaster* Acps are not found in distantly related species (Mueller et al. 2005; Wagstaff and Begun 2005a; Haerty et al. 2007) (although in some cases this may be due to sequence divergence), and many Acps from other species of *Drosophila* are similarly lineage specific (Begun and Lindfors 2005; Begun et al. 2006) . The population samples that we sequenced in this study harbored a number of putative loss of function alleles at multiple loci (Table 5.4), that may represent loci becoming pseudogenized. In two cases (CG31681, CG32383), a single allele was sequenced with a premature stop codon. The low frequencies of these alleles may be consistent with mutation-selection balance. However, in the case of CG14642, a female-expressed protease, 5 out of 16 *D. melanogaster* alleles carried single base-pair frameshifts, due to at least three independent mutational events. The frequency of apparent LOF alleles at this locus is too high to be explained solely by mutation-selection balance, and it is tempting to posit that CG14642 is in transit towards becoming a pseudogene. Finally, the predicted protease inhibitor CG9334 may have recently become a pseudogene in *D. simulans*. Every allele sampled bore multiple frameshift mutations and indels, although work is still underway to determine whether paralogous copies exist in the *D. simulans* genome.

Discussion

A variety of hypotheses have been proposed to explain the observation that reproductive tract proteins in Drosophila and other organisms tend to evolve rapidly and adaptively. Male-female co-evolution, sperm competition, and host-pathogen interactions are among the leading proposals (Swanson and Vacquier 2002; Clark,

Table 5.4 Putative loss-of-function (LOF) alleles amongst 37 proteolysis regulators and 3 targets of proteolysis

Gene	# LOF alleles	Type	Frequency
CG31681 (*D. melanogaster*)	1	Premature stop	1/14
CG32383 (*D. melanogaster*)	1	Premature stop	1/14
CG14642 (*D. melanogaster*)	3	Frameshift	3/16, 1/16, 1/16

Aagaard, and Swanson 2006; Panhuis, Clark, and Swanson 2006), but it has proven difficult to distinguish between these potential mechanisms.

Some of the strongest evidence for co-evolution between male and female reproductive proteins in animals comes from abalone. In abalone, sperm-egg fusion is mediated by the sperm protein lysin, and VERL, its receptor on the egg. Lysin evolves very rapidly between species, under the influence of positive selection (Swanson, Aquadro, and Vacquier 2001), and functional studies have demonstrated a high degree of species specificity in lysin-VERL interactions (Vacquier and Lee 1993). VERL largely consists of a tandemly repeated subunit, and earlier studies showed that a number of VERL's repeats undergo concerted evolution. A model was proposed whereby positive selection on lysin occurs as a response to rapid changes in repeat sequence in VERL. A recent report (Galindo, Vacquier, and Swanson 2003), however, shows that the first repeat of VERL also undergoes rapid, adaptive evolution, suggesting a tight coupling of selection at lysin and at VERL. Intriguingly, polymorphism data within abalone populations suggests linkage disequilibrium between the *lysin* and *VERL* loci, a strong expectation under mate choice models (Clark, Springer, and Swanson, in preparation).

Evidence for co-evolution between male and female reproductive proteins in most other organisms, however, is weaker, largely owing to a lack of knowledge concerning the identities of interacting partners (e.g., Chapter 1, but see Nasrallah 2002 for an exception). In *D. melanogaster*, a model organism for which a wealth of information regarding reproductive protein function and evolution is available (Panhuis, Clark, and Swanson 2006; Ravi Ram and Wolfner 2007), data regarding co-evolution is growing. Positive selection has been documented on several genes encoding male (Table 1.2) and female (Swanson et al. 2004; Panhuis and Swanson 2006; Kelleher, Swanson, and Markow 2007; Lawniczak and Begun 2007)

reproductive tract proteins. Moreover, comparisons of evolutionary rates for male and female reproductive tract proteins in species with different mating rates are consistent with an important role for post-copulatory sexual selection in the evolution of these molecules.

In order to further evaluate the role of post-copulatory sexual selection, and male-female co-evolution in particular, in driving the evolution of reproductive tract proteins, we conducted molecular population genetic surveys for 40 genes encoding potentially interacting proteins – proteolysis regulators and targets of proteolysis – expressed in the male and female reproductive tracts of *D. melanogaster*. We found strong evidence in favor of male-female molecular co-evolution, with similar levels of positive selection on male- and female- reproductive tract proteins, and an excess of linkage disequilibrium between male- and female- expressed genes in *D. simulans*. In the future, we intend to perform additional inferences of selection based on the site-frequency spectrum (Nielsen et al. 2005); careful attention to demographic history will be required in these analyses, given the evidence for non-equilibrium population dynamics in *D. simulans* (Figure 5.1; a negative Tajima's D in *D. simulans* may be suggestive of recent growth). In addition, further LD-based tests will be performed, including analyses polarizing changes as ancestral or derived, and Lewontin's sign test (Lewontin 1995).

Haerty et al. (2007) found, for 679 female reproductive tract genes, a lower average rate of protein evolution, and lower levels of positive selection, than male reproductive tract genes. Our study has somewhat different findings: Mean γ for 13 female reproductive tract genes was higher than that for 26 male reproductive tract genes (this difference was not significant), although a lower proportion of individual female reproductive tract genes showed strong evidence for positive selection (the difference was again not significant). Taken together, our data suggest that the

strength of selection on male and female reproductive tract proteolysis regulators and targets does not differ greatly, in contrast to the findings of Haerty et al. (2007). We suggest two possible explanations for this disparity. First, the proteins encoded by the set of genes examined in this study may include a higher proportion of true interactors, given their predicted biochemical functions; the loci examined by Haerty et al. (2007) were chosen solely on the basis of location of expression. Second, while we used within-species polymorphism analyses that are sensitive to relatively recent signatures of selection, Haerty et al. (2007) used between-species divergence analyses that detect recurrent selection on the same codons, in the same genes, across deep evolutionary time. The disparity between the two studies might be explained if the set of rapidly evolving male reproductive tract genes (and codons therein) is more consistent across taxa than is the set of rapidly evolving female reproductive tract genes.

Our data additionally suggest that immune interactions may play an important role in reproductive tract protein evolution, since several genes showing evidence for positive selection have documented roles in regulating the immune response. We note that we cannot exclude the possibility that such genes play multiple, independent roles in immunity and in reproduction. Tests of post-mating responses (ovulin cleavage, remating propensity, egg-laying) in females knocked-down for one positively selected female reproductive tract protease, CG3066, have failed to reveal any reproductive function for this gene (Wong, Sirot, Guise, and Wolfner, unpublished data), although not all post-mating responses have been examined. Given these findings, we suggest that the hypothesis that host-pathogen interactions contribute to positive selection in the reproductive tract warrants further consideration.

The current results do not allow us to distinguish between different forms of male-female co-evolution, e.g., sexual conflict or female cryptic choice, in driving the evolution of particular loci; we suspect that such inferences cannot be made from

population genetic data alone. Moreover, we cannot unequivocally exclude other hypotheses concerning the evolution of reproductive tract proteins. Male-male competition, for example, is likely to be an important force in reproductive protein evolution. In our data, a slight elevation of LD in male-male comparisons over female-female comparisons may be suggestive of male-male interactions. However, this elevation may be due to either (or both of) interactions between ejaculates (sperm competition) or within the ejaculate of a single male, and as such is inconclusive at this point. Furthermore, our data do not allow us address such hypotheses as reinforcement (which would require population genetic data from incipient species) or avoidance of inbreeding.

Our work provides additional evidence in favor of male-female co-evolution in driving the evolution of reproductive tract proteins in Drosophila, and further supports the idea that immune interactions play an important role. Genetic and biochemical tests of interactions between promising candidate genes, such as CG3074 and Acp36DE, will help to confirm our inferences.

REFERENCES

Adams, E. M., and M. F. Wolfner. 2007. Seminal proteins but not sperm induce morphological changes in the *Drosophila melanogaster* female reproductive tract during sperm storage. J Insect Physiol **53**:319-331.

Aguadé, M. 1998. Different forces drive the evolution of the Acp26Aa and Acp26Ab accessory gland genes in the *Drosophila melanogaster* species complex. Genetics **150**:1079-1089.

Aguadé, M. 1999. Positive selection drives the evolution of the Acp29AB accessory gland protein in Drosophila. Genetics **152**:543-551.

Aguadé, M., N. Miyashita, and C. H. Langley. 1992. Polymorphism and divergence in the Mst26A male accessory gland gene region in Drosophila. Genetics **132**:755-770.

Allen, A. K., and A. C. Spradling. 2008. The Sf1-related nuclear hormone receptor Hr39 regulates Drosophila female reproductive tract development and function. Development **135**:311-321.

Andersson, M. 1994. Sexual Selection. Princeton University Press, Princeton, NJ.

Andolfatto, P. 2005. Adaptive evolution of non-coding DNA in Drosophila. Nature **437**:1149-1152.

Asgari, S., G. Zhang, R. Zareie, and O. Schmidt. 2003. A serine proteinase homolog venom protein from an endoparasitoid wasp inhibits melanization of the host hemolymph. Insect Biochem Mol Biol **33**:1017-1024.

Ball, M. A., and G. A. Parker. 2003. Sperm competition games: sperm selection by females. J Theor Biol **224**:27-42.

Barrier, M., C. D. Bustamante, J. Yu, and M. D. Purugganan. 2003. Selection on rapidly evolving proteins in the Arabidopsis genome. Genetics **163**:723-733.

Begun, D. J., A. K. Holloway, K. Stevens, L. W. Hillier, Y. P. Poh, M. W. Hahn, P. M. Nista, C. D. Jones, A. D. Kern, C. N. Dewey, L. Pachter, E. Myers, and C. H. Langley. 2007. Population genomics: whole-genome analysis of polymorphism and divergence in *Drosophila simulans*. PLoS Biol **5**:e310.

Begun, D. J., and H. A. Lindfors. 2005. Rapid evolution of genomic Acp complement in the melanogaster subgroup of Drosophila. Mol Biol Evol **22**:2010-2021.

Begun, D. J., H. A. Lindfors, M. E. Thompson, and A. K. Holloway. 2006. Recently evolved genes identified from *Drosophila yakuba* and *D. erecta* accessory gland expressed sequence tags. Genetics **172**:1675-1681.

Begun, D. J., P. Whitley, B. L. Todd, H. M. Waldrip-Dail, and A. G. Clark. 2000. Molecular population genetics of male accessory gland proteins in Drosophila. Genetics **156**:1879-1888.

Bustamante, C. D., R. Nielsen, S. A. Sawyer, K. M. Olsen, M. D. Purugganan, and D. L. Hartl. 2002. The cost of inbreeding in Arabidopsis. Nature **416**:531-534.

Cameron, E., T. Day, and L. Rowe. 2007. Sperm competition and the evolution of ejaculate composition. Am Nat **169**:e158-e172.

Chapman, T., L. F. Liddle, J. M. Kalb, M. F. Wolfner, and L. Partridge. 1995. Cost of mating in *Drosophila melanogaster* females is mediated by male accessory gland products. Nature **373**:241-244.

Chenoweth, S. F., H. D. Rundle, and M. W. Blows. 2008. Genetic constraints and the evolution of display trait sexual dimorphism by natural and sexual selection. Am Nat **171**:22-34.

Cirera, S., and M. Aguadé. 1997. Evolutionary history of the sex-peptide (Acp70A) gene region in *Drosophila melanogaster*. Genetics **147**:189-197.

Civetta, A., and A. G. Clark. 2000. Chromosomal effects on male and female components of sperm precedence in Drosophila. Genet Res **75**:143-151.

Civetta, A., and R. S. Singh. 1995. High divergence of reproductive tract proteins and their association with postzygotic reproductive isolation in *Drosophila melanogaster* and *Drosophila virilis* group species. J Mol Evol **41**:1085-1095.

Clark, A. G., M. Aguadé, T. Prout, L. G. Harshman, and C. H. Langley. 1995. Variation in sperm displacement and its association with accessory gland protein loci in *Drosophila melanogaster*. Genetics **139**:189-201.

Clark, A. G., D. J. Begun, and T. Prout. 1999. Female x male interactions in Drosophila sperm competition. Science **283**:217-220.

Clark, N. L., J. E. Aagaard, and W. J. Swanson. 2006. Evolution of reproductive proteins from animals and plants. Reproduction **131**:11-22.

Darwin, C. 1871. The Descent of Man, and Selection in Relation to Sex. John Murray, London.

Dimmic, M. W., M. J. Hubisz, C. D. Bustamante, and R. Nielsen. 2005. Detecting coevolving amino acid sites using Bayesian mutational mapping. Bioinformatics **21 Suppl 1**:i126-135.

Domanitskaya, E. V., H. Liu, S. Chen, and E. Kubli. 2007. The hydroxyproline motif of male sex peptide elicits the innate immune response in Drosophila females. FEBS J **274**:5659-5668.

Eberhard, W. G. 1996. Female control: Sexual selection by cryptic female choice. Princeton University Press, Princeton, N. J.

Fiumera, A. C., B. L. Dumont, and A. G. Clark. 2007. Associations between sperm competition and natural variation in male reproductive genes on the third chromosome of *Drosophila melanogaster*. Genetics **176**:1245-1260.

Fiumera, A. C., B. L. Dumont, and A. G. Clark. 2005. Sperm competitive ability in *Drosophila melanogaster* associated with variation in male reproductive proteins. Genetics **169**:243-257.

Galindo, B. E., V. D. Vacquier, and W. J. Swanson. 2003. Positive selection in the egg receptor for abalone sperm lysin. Proc Natl Acad Sci U S A **100**:4639-4643.

Gupta, S., Y. Wang, and H. Jiang. 2005. *Manduca sexta* prophenoloxidase (proPO) activation requires proPO-activating proteinase (PAP) and serine proteinase homologs (SPHs) simultaneously. Insect Biochem Mol Biol **35**:241-248.

Haerty, W., S. Jagadeeshan, R. J. Kulathinal, A. Wong, K. Ravi Ram, L. K. Sirot, L. Levesque, C. G. Artieri, M. F. Wolfner, A. Civetta, and R. S. Singh. 2007. Evolution in the fast lane: rapidly evolving sex-related genes in Drosophila. Genetics **177**:1321-1335.

Heifetz, Y., L. N. Vandenberg, H. I. Cohn, and M. F. Wolfner. 2005. Two cleavage products of the Drosophila accessory gland protein ovulin can independently induce ovulation. Proc Natl Acad Sci U S A **102**:743-748.

Higgie, M., and M. W. Blows. 2008. The Evolution of Reproductive Character Displacement Conflicts with How Sexual Selection Operates within a Species. Evolution Int J Org Evolution.

Kelleher, E. S., W. J. Swanson, and T. A. Markow. 2007. Gene duplication and adaptive evolution of digestive proteases in *Drosophila arizonae* female reproductive tracts. PLoS Genet **3**:1541-1549.

Keller, L., and H. Reeve. 1995. Why do females mate with multiple males? The sexually selected sperm hypothesis. Advances in the study of behavior **24**:291-315.

Kelly, J. K. 1997. A test of neutrality based on interlocus associations. Genetics **146**:1197-1206.

Kirkpatrick, M. 1982. Sexual selection and the evolution of female choice. Evolution **36**:1-12.

Kwon, T. H., M. S. Kim, H. W. Choi, C. H. Joo, M. Y. Cho, and B. L. Lee. 2000. A masquerade-like serine proteinase homologue is necessary for phenoloxidase activity in the coleopteran insect, Holotrichia diomphalia larvae. Eur J Biochem **267**:6188-6196.

Lawniczak, M. K., A. I. Barnes, J. R. Linklater, J. M. Boone, S. Wigby, and T. Chapman. 2007. Mating and immunity in invertebrates. Trends Ecol Evol **22**:48-55.

Lawniczak, M. K., and D. J. Begun. 2007. Molecular Population Genetics of Female-expressed Mating-induced Serine Proteases in *Drosophila melanogaster*. Mol Biol Evol.

Lawniczak, M. K., and D. J. Begun. 2004. A genome-wide analysis of courting and mating responses in *Drosophila melanogaster* females. Genome **47**:900-910.

Lee, K. Y., R. Zhang, M. S. Kim, J. W. Park, H. Y. Park, S. Kawabata, and B. L. Lee. 2002. A zymogen form of masquerade-like serine proteinase homologue is cleaved during pro-phenoloxidase activation by Ca2+ in coleopteran and *Tenebrio molitor* larvae. Eur J Biochem **269**:4375-4383.

Lewontin, R. C. 1995. The detection of linkage disequilibrium in molecular sequence data. Genetics **140**:377-388.

Lung, O., L. Kuo, and M. F. Wolfner. 2001. Drosophila males transfer antibacterial proteins from their accessory gland and ejaculatory duct to their mates. J Insect Physiol **47**:617-622.

Mack, P. D., A. Kapelnikov, Y. Heifetz, and M. Bender. 2006. Mating-responsive genes in reproductive tissues of female *Drosophila melanogaster*. Proc Natl Acad Sci U S A **103**:10358-10363.

McDonald, J. H., and M. Kreitman. 1991. Adaptive protein evolution at the Adh locus in Drosophila. Nature **351**:652-654.

McGraw, L. A., G. Gibson, A. G. Clark, and M. F. Wolfner. 2004. Genes regulated by mating, sperm, or seminal proteins in mated female *Drosophila melanogaster*. Curr Biol **14**:1509-1514.

Miller, G. T., and S. Pitnick. 2002. Sperm-female coevolution in Drosophila. Science **298**:1230-1233.

Mueller, J. L., J. L. Page, and M. F. Wolfner. 2007. An ectopic expression screen reveals the protective and toxic effects of Drosophila seminal fluid proteins. Genetics **175**:777-783.

Mueller, J. L., K. R. Ram, L. A. McGraw, M. C. Bloch Qazi, E. D. Siggia, A. G. Clark, C. F. Aquadro, and M. F. Wolfner. 2005. Cross-species comparison of Drosophila male accessory gland protein genes. Genetics **171**:131-143.

Mueller, J. L., D. R. Ripoll, C. F. Aquadro, and M. F. Wolfner. 2004. Comparative structural modeling and inference of conserved protein classes in Drosophila seminal fluid. Proc Natl Acad Sci U S A **101**:13542-13547.

Nasrallah, J. B. 2002. Recognition and rejection of self in plant reproduction. Science **296**:305-308.

Nielsen, R., S. Williamson, Y. Kim, M. J. Hubisz, A. G. Clark, and C. Bustamante. 2005. Genomic scans for selective sweeps using SNP data. Genome Res **15**:1566-1575.

Nolte, V., and C. Schlotterer. 2008. African *Drosophila melanogaster* and D. simulans populations have similar levels of sequence variability, suggesting comparable effective population sizes. Genetics **178**:405-412.

Panhuis, T. M., N. L. Clark, and W. J. Swanson. 2006. Rapid evolution of reproductive proteins in abalone and Drosophila. Philos Trans R Soc Lond B Biol Sci **361**:261-268.

Panhuis, T. M., and W. J. Swanson. 2006. Molecular evolution and population genetic analysis of candidate female reproductive genes in Drosophila. Genetics **173**:2039-2047.

Park, M., and M. F. Wolfner. 1995. Male and female cooperate in the prohormone-like processing of a *Drosophila melanogaster* seminal fluid protein. Dev Biol **171**:694-702.

Peng, J., S. Chen, S. Busser, H. Liu, T. Honegger, and E. Kubli. 2005. Gradual release of sperm bound sex-peptide controls female postmating behavior in Drosophila. Curr Biol **15**:207-213.

Peng, J., P. Zipperlen, and E. Kubli. 2005. Drosophila sex-peptide stimulates female innate immune system after mating via the Toll and Imd pathways. Curr Biol **15**:1690-1694.

Pitnick, S., G. T. Miller, K. Schneider, and T. A. Markow. 2003. Ejaculate-female coevolution in *Drosophila mojavensis*. Proc Biol Sci **270**:1507-1512.

Pool, J. E., and C. F. Aquadro. 2006. History and structure of sub-Saharan populations of *Drosophila melanogaster*. Genetics **174**:915-929.

Ravi Ram, K., L. K. Sirot, and M. F. Wolfner. 2006. Predicted seminal astacin-like protease is required for processing of reproductive proteins in *Drosophila melanogaster*. Proc Natl Acad Sci U S A **103**:18674-18679.

Ravi Ram, K., and M. F. Wolfner. 2007. Seminal influences: *Drosophila* Acps and the molecular interplay between males and females during reproduction. Integrative and Comparative Biology. **47**: 427-445.

Sackton, T. B., B. P. Lazzaro, T. A. Schlenke, J. D. Evans, D. Hultmark, and A. G. Clark. 2007. Dynamic evolution of the innate immune system in Drosophila. Nat Genet **39**:1461-1468.

Swanson, W. J., C. F. Aquadro, and V. D. Vacquier. 2001. Polymorphism in abalone fertilization proteins is consistent with the neutral evolution of the egg's receptor for lysin (VERL) and positive darwinian selection of sperm lysin. Mol Biol Evol **18**:376-383.

Swanson, W. J., A. G. Clark, H. M. Waldrip-Dail, M. F. Wolfner, and C. F. Aquadro. 2001. Evolutionary EST analysis identifies rapidly evolving male reproductive proteins in Drosophila. Proc Natl Acad Sci U S A **98**:7375-7379.

Swanson, W. J., and V. D. Vacquier. 2002. The rapid evolution of reproductive proteins. Nat Rev Genet **3**:137-144.

Swanson, W. J., A. Wong, M. F. Wolfner, and C. F. Aquadro. 2004. Evolutionary expressed sequence tag analysis of Drosophila female reproductive tracts identifies genes subjected to positive selection. Genetics **168**:1457-1465.

Thornton, K. 2003. Libsequence: a C++ class library for evolutionary genetic analysis. Bioinformatics **19**:2325-2327.

Tsaur, S. C., C. T. Ting, and C. I. Wu. 1998. Positive selection driving the evolution of a gene of male reproduction, Acp26Aa, of Drosophila: II. Divergence versus polymorphism. Mol Biol Evol **15**:1040-1046.

Tsaur, S. C., and C. I. Wu. 1997. Positive selection and the molecular evolution of a gene of male reproduction, Acp26Aa of Drosophila. Mol Biol Evol **14**:544-549.

Vacquier, V. D., and Y. H. Lee. 1993. Abalone sperm lysin: unusual mode of evolution of a gamete recognition protein. Zygote **1**:181-196.

Wagstaff, B. J., and D. J. Begun. 2005a. Comparative genomics of accessory gland protein genes in *Drosophila melanogaster* and *D. pseudoobscura*. Mol Biol Evol **22**:818-832.

Wagstaff, B. J., and D. J. Begun. 2005b. Molecular Population Genetics of Accessory Gland Protein Genes and Testis-expressed Genes in *Drosophila mojavensis* and *D. arizonae*. Genetics.

Wagstaff, B. J., and D. J. Begun. 2007. Adaptive evolution of recently duplicated accessory gland protein genes in desert Drosophila. Genetics **177**:1023-1030.

Wigby, S., and T. Chapman. 2005. Sex peptide causes mating costs in female *Drosophila melanogaster*. Curr Biol **15**:316-321.

Wigby, S., E. V. Domanitskaya, Y. Choffat, E. Kubli, and T. Chapman. 2008. The effect of mating on immunity can be masked by experimental piercing in female *Drosophila melanogaster*. J Insect Physiol **54**:414-420.

Wong, A., M. C. Turchin, M. F. Wolfner, and C. F. Aquadro. 2008. Evidence for positive selection on *Drosophila melanogaster* seminal fluid protease homologs. Mol Biol Evol **25**:497-506.

Yapici, N., Y. J. Kim, C. Ribeiro, and B. J. Dickson. 2008. A receptor that mediates the post-mating switch in Drosophila reproductive behaviour. Nature **451**:33-37.

CHAPTER 6

PHYLOGENETIC INCONGRUENCE IN THE *DROSOPHILA MELANOGASTER* SPECIES GROUP[1]

Introduction

Drosophila melanogaster and its relatives have been used extensively in studies of genetic and morphological variation within and between species. For example, inferences concerning the relative roles of drift, purifying selection, and positive selection in shaping patterns of genetic variation in *D. melanogaster* often benefit from comparisons to the closely related species *D. simulans* and *D. yakuba* (e.g., McDonald and Kreitman 1991). Similarly, comparative morphologists have used *D. melanogaster* and its relatives to study the evolution of a number of traits, e.g., genital morphology (Kopp and True 2002a) and pigmentation (Wittkopp, True, and Carroll 2002; Prud'homme et al. 2006).

Opportunities for, and interest in, using the genus Drosophila in comparative biology is likely to grow in the near future. The availability of complete genome sequences for twelve members of the genus Drosophila (http://species.flybase.net), as well as for several other dipterans, promises to facilitate genome scale studies of molecular evolution. These comparative data will allow for the detection of functionally important genomic regions, as indicated by high levels of conservation or by the signature of positive, diversifying selection. Moreover, the application of

[1] This chapter was published previously as: Wong A, Jensen JD, Pool JE, Aquadro CF. 2007. Phylogenetic incongruence in the *melanogaster* species group. *Mol Phy Evol* 43(3): 1138-50. Jeff Jensen performed some of the phylogenetic tree reconstruction described herein, helped to compile genes for analysis, and provided conceptual and intellectual input. John Pool sequenced one locus used in this study (seq211), and provided conceptual and intellectual input. I performed the remaining sequencing and analyses, and co-wrote the manuscript with CFA. Copyright permissions for theses are automatically granted by the journal.

genetic and transgenic techniques developed in *D. melanogaster* to other species will facilitate studies of evolution and development.

Different levels of taxonomic organization have proven useful for comparisons of different traits of interest. Rapidly evolving characters, such as genital morphology, necessitate the use of closely related taxa (e.g., Kopp and True 2002a). Over longer taxonomic distances, it may become difficult to distinguish the ancestral from the derived state, because all extant taxa will be highly derived. Moreover, the likelihood of observing homoplasies (independent mutational events leading to a shared character state) increases with greater evolutionary time. The study of slowly evolving characters, by contrast, requires the use of more distantly related species, such that sufficient time has elapsed in order to observe evolutionary change.

With respect to *D. melanogaster*, we expect that comparisons within the *melanogaster* subgroup and group will be particularly relevant to many comparative studies (Figure 6.1), particularly in comparative genomics. With greater phylogenetic distance, synonymous sites become saturated, undermining the utility of dN/dS based measures of molecular evolution. In comparisons between the fully sequenced genomes of *D. melanogaster* and *D. pseudoobscura*, for example, enough synonymous sites have sustained multiple hits to substantially reduce the power and reliability of the dN/dS ratio (Richards et al. 2005). The so-called "oriental" subgroups (*takahashii*, *eugracilis*, *elegans*, *suzukii*, *ficusphila*, *rhopaloa*), which are thought to be intermediate in divergence between *D. melanogaster* and *D. pseudoobscura* (Lemeunier, David, and Tsacas 1986), may therefore be of particular use, since synonymous sites are typically not saturated (e.g., Swanson et al., 2004; Malik and Henikoff, 2005). Moreover, the species comprising the oriental subgroups display an impressive array of morphological diversity (e.g., Kopp and True 2002a; Prud'homme et al. 2006).

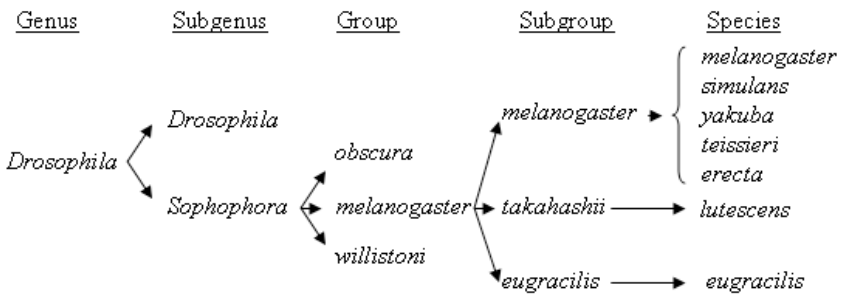


Figure 6.1 Taxonomic subdivisions in the genus Drosophila. Only species and subgroups represented in this study are listed; other groups and subgenuses are indicated for illustrative purposes only.

Most statistical methods used in comparative genomics and comparative morphology require explicit use of a phylogeny of the taxa under consideration. For example, PAML, a software package used frequently for detecting positive selection at the codon level, requires specification of a tree or trees upon which evolutionary parameters are estimated (Yang et al. 2000). Phylogenetic considerations are crucial; for example, it is only through use of a phylogeny that one can distinguish between shared genealogy and convergent evolution as explanations for a shared character state. A robust phylogeny of the *Drosophila melanogaster* species group will therefore prove important for future comparative work.

Despite numerous attempts to infer phylogenies within the *Drosophila melanogaster* species group, several relationships have proven difficult to resolve. Within the *melanogaster* subgroup, three pairs of sibling species (or species complexes) are well established: *melanogaster*/*simulans* (and associated *simulans* complex species), *erecta*/*orena*, and *teissieri*/*yakuba* (and *D. santomea*). It is thought that the three species complexes of the *melanogaster* subgroup diverged between 6 and 15 million years ago (Lachaise et al. 1988). The relationships among these species pairs have proven controversial (Figure 6.2), although recent molecular studies appear to converge on a single topology (Kopp and True 2002b; Ko, David, and Akashi 2003). LaChaise et al. (1988), on the basis of biogeographic considerations, places the *erecta*/*orena* clade basal within the subgroup (this configuration is denoted Topology I by Ko et al. 2003, whose nomenclature we follow here). Jeffs et al. (1994) and Russo et al. (1995) support this hypothesis using nuclear gene sequence data. Several other studies find evidence for a closer relationship between the *teissieri*/*yakuba* and *erecta*/*orena* species pairs (Topology II; Gailey et al. 2000; Arhontaki et al. 2002; Ko, David, and Akashi 2003). Finally, one study places *D.*

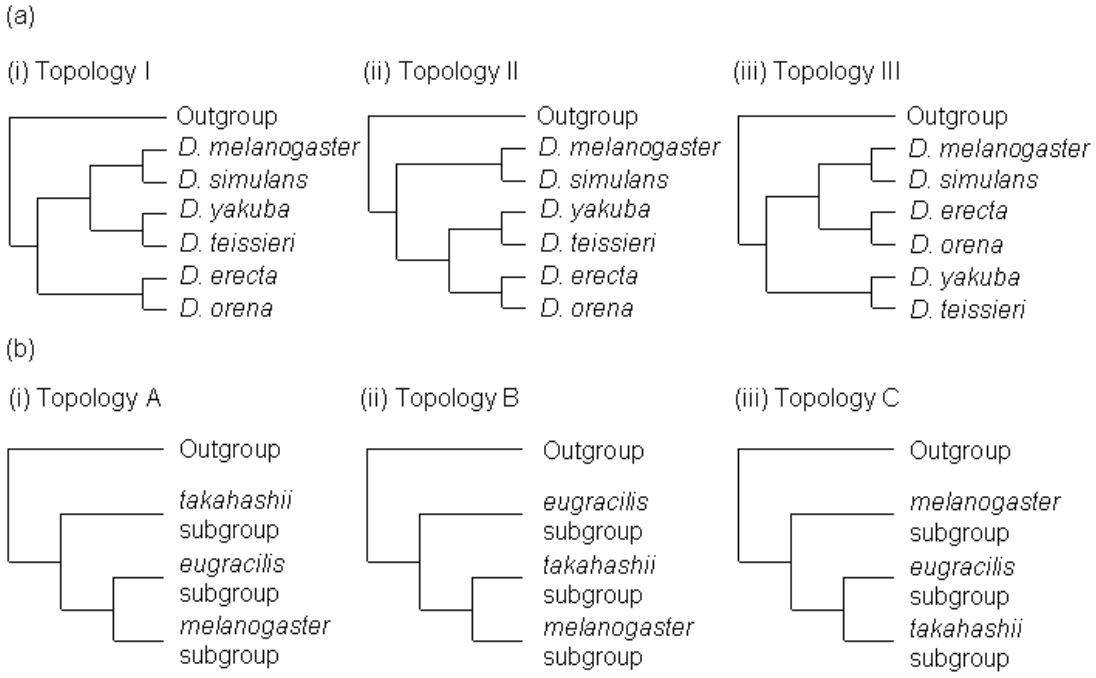


Figure 6.2 (A): Possible tree topologies of the *melanogaster* subgroup (B): Possible tree topologies of the *melanogaster* species group.

erecta and *D. orena* closest to the *melanogaster*/*simulans* complex (Topology III; Schlotterer et al. 1994).

Relationships between the *melanogaster* subgroup and the oriental subgroups have also been difficult to resolve (Figure 6.2). Here, we focus on the branching orders of the *eugracilis*, *takahashii*, and *melanogaster* subgroups, which likely diverged between 15 and 30 million years ago (Lachaise et al. 1988). Analyses of several nuclear genes place the *takahashii* subgroup basal within the species group, with strong bootstrap support (we will call this Topology A; Ko et al., 2003). Other studies, with similarly strong support, find a basal position for the *eugracilis* subgroup (Topology B; Kopp and True 2002b; Yang et al. 2004). A third topology, according to which the *eugracilis* and *takahashii* subgroups are more closely related to each other than either is to the *melanogaster* subgroup (Topology C), is supported by mtDNA (Kastanis et al. 2003).

Although previous studies have used multiple loci to infer different phylogenies within the *melanogaster* species group and subgroup, none has explicitly addressed the issue of incongruence between loci. It is unclear whether apparent disagreements between loci are statistically robust, and the underlying causes of incongruence have not been addressed. Here, we use twelve nuclear loci, representing eleven protein coding genes (of which ten are autosomal and one X-linked in *D. melanogaster*) and one non-coding region (X-linked in *D. melanogaster*), to test for phylogenetic incongruence and to investigate its causes. Within the *melanogaster* subgroup, we use sequences from *D. melanogaster*, *D. simulans*, *D. yakuba*, *D. teissieri*, and *D. erecta*. *D. eugracilis* and *D. lutescens* serve as representatives of the *eugracilis* and *takahashii* subgroups, respectively. We use sequences from *D. pseudoobscura* and *D. ananassae* as outgroups.

Using maximum parsimony, maximum likelihood, and Bayesian phylogenetic reconstruction methods, we find strong support for Topology II (*D. yakuba*/*D. teissieri* + *D. erecta*/*D. orena*) within the *melanogaster* subgroup. Relationships among the *melanogaster*, *eugracilis*, and *takahashii* subgroups remain equivocal, however, with different loci supporting different tree topologies. Using the likelihood heterogeneity test (LHT) of Huelsenbeck and Bull (1996), we find statistically robust evidence for topological incongruence between loci, which we argue cannot be attributed to a variety of potential confounding factors. In light of the difficulty in resolving relationships between these three subgroups, in this and other studies, we propose that these lineages may have speciated rapidly from a common, polymorphic ancestor, such that lineage sorting resulted in incongruent trees for different gene regions (Pamilo and Nei 1988). Interestingly, we find evidence for intralocus recombination in the common ancestor of the *melanogaster*, *eugracilis*, and *takahashii* subgroups, and in the common ancestor of the *melanogaster* subgroup. We discuss the possible implications of such complex histories for inferences of tree topology, substitution rates, and positive selection.

Materials and Methods

Drosophila *Strains, DNA Sequences, and Sequence Alignment*

Most sequences used in this study have been previously published, and were obtained from public databases (Table 6.1). Several additional sequences were collected for this study, from the following strains kindly donated by Andrew Clark (Cornell University): *D. erecta* (S-18; originally from the Ashburner laboratory), *D. eugracilis* (Tucson Drosophila Stock Center 451.3), *D. lutescens* (271.1), *D. teissieri* (257.0), and *D. yakuba* (261.0). *D. simulans* sequences were from an Australian isofemale line collected in December, 1997 by Ary Hoffmann. Partial coding sequences

Table 6.1 Loci used in chapter 6. Genomic location is given as the chromosome arm and cytological band in *D. melanogaster*. Tree length is the total tree length in expected substitutions per site, from the maximum likelihood tree.

Locus	Coding/Non-coding	Genomic location in *D. melanogaster*[a]	Length	Reference	Tree length[b]
Adh	coding	2L (35B3)	834	Ko et al. (2003)	0.57
Adhr	coding	2L (35B3)	875	Ko et al. (2003)	0.88
ry	coding	3R (87D9)	4098	Ko et al. (2003)	0.99
Gld	coding	3R (84D3)	1549	Ko et al. (2003)	0.85
mitch	coding	3 (87D5)	699	Goldberg et al. (unpublished)	1.79
hb	coding	3R (85A5)	534	Schawaroch (2000)	1.61
CG3066	coding	3R (84D14-E1)	872	This study	1.42
CG4928	coding	X (15C1-4)	1536	This study	0.43
CG7415	coding	3R (84F13)	788	This study	1.02
CG9336	coding	2L (38F3)	378	This study	0.67
seq211	non-coding	X (3C5)	2859	This study	1.25
Iris	coding	2L (21F1)	1620	Malik and Henikoff (2005)	2.53

for CG3066, CG7415, CG4928 were used by Swanson et al. (2004) for inferences of positive selection. Sequences from additional species for these genes have been deposited in GenBank under the following accession numbers: DQ907915, DQ907916, and DQ907923. The full coding sequence of *mitch* was obtained from GenBank for all species except *D. ananassae*. Sequence for *D. ananassae* was obtained from the public sequencing effort (http://species.flybase.net). Sequences for CG9336 and the non-coding locus seq211 have not been previously published, and have been deposited in GenBank under accession numbers DQ907917- DQ907922, and DQ907924- DQ907929. Sequences for *Adh*, *Adhr*, *Gld*, and *ry* were obtained from Ko et al. (2003), with the exception of sequences from *D. ananassae*, which were obtained from the public sequencing effort. *hunchback* (*hb*) sequences are from (Schawaroch 2002), and *Iris* sequences are from Malik and Henikoff (2005). Sequence alignments for coding regions were performed using the ClustalW algorithm, as implemented in MegAlign (DNASTAR, Inc.), and were modified by eye to maximize amino acid identity. The non-coding locus seq211 was aligned using MAVID (Bray and Pachter 2004).

Tests for Saturation and Base Compositional Bias

We tested for substitutional saturation, in order to assess the potential effects of homoplasy on phylogenetic inferences. Following Engstrom et al. (2004), for each locus, the uncorrected distance (*p*) between each pair of species was plotted against the maximum likelihood corrected distance (*ML*). A positive relationship is expected for unsaturated data, while saturated data plateau at higher levels of divergence. To identify such a plateau, we fitted a second order polynomial to each of the saturation plots using the statistical package JMP IN 5.1 (Duxbury). We then identified the maximum of the regression line, which represents the point at which a positive

relationship no longer exists between *p* and *ML*. Data points to the right of the maximum suffer from saturation, raising homoplasy as a concern.

For each locus, chi-squared tests for base frequency equilibrium across all species (including outgroups) were performed using PAUP*4.0b10 (Swofford 2002).

Phylogenetic Inference

Maximum parsimony and maximum likelihood analyses were performed using PAUP*4.0b10 (Swofford 2002). For individual locus analyses and for the concatenated alignment, maximum likelihood analyses were performed under the general time reversible model of nucleotide substitution, with gamma distributed rates, allowing for invariant sites (GTR+G+I; Felsenstein 1981; Yang 1994). MrBayes 3.0b4 was used for Bayesian phylogeny estimation (Huelsenbeck and Ronquist 2001; Ronquist and Huelsenbeck 2003). We again used the GTR+G+I model of nucleotide substitution. In single locus analyses, four Markov chains were run for 100,000 generations of burn-in, followed by 500,000 generations for topology and parameter estimation. For the concatenated data set, four chains were allowed to run for 2,000,000 generations, following 500,000 generations of burn-in.

Interior branch length tests

At each locus, we used likelihood ratio tests (LRT) as implemented in PAUP*4.0b10 (Swofford 2002) to test for zero branch lengths around two nodes: the node connecting *D. eugracilis*, *D. lutescens*, and the *melanogaster* subgroup, and the node connecting *D. erecta*, the *D. simulans*/*D. melanogaster* species pair, and the *D. yakuba*/*D. teissieri* species pair. In this LRT, the null hypothesis (H_0) is that the branch in question has zero length (i.e., that the relevant node is a molecular polytomy). The alternative hypothesis (H_A) states that the branch has a positive

length. The LRT test statistic, $2[\ln(L_{H0}) - \ln(L_{HA})]$, where L_{H0} and L_{HA} represent the likelihoods of H_0 and H_A respectively, follows a 50:50 mixture distribution of the χ^2 with 0 degrees of freedom and the χ^2 with 1 degree of freedom (Goldman and Whelan 2000; Slowinski 2001).

Statistical tests of incongruence

We performed two tests of incongruence. First, we applied the incongruence length difference (ILD) test (Farris et al. 1995), as implemented under the partition homogeneity test in PAUP*4.0b10 (Swofford 2002). This commonly used test compares the length of the most parsimonious tree under user defined data partitions (here, different loci) to the length of the most parsimonious tree for the combined data. The null distribution is obtained by creating new partitions of the same size as the user defined partitions at random from the original dataset. One thousand bootstrap replicates were used for the null distribution.

Since the ILD test may reject the null hypothesis of congruence for reasons other than topological incongruence (e.g., Darlu and Lecointre 2002), and does not readily allow for localization of incongruence to specific nodes, we implemented the LHT of Huelsenbeck and Bull (1996). The null hypothesis (H_0) of the LHT states that the same topology underlies all data partitions (in this case, different loci), while the alternative hypothesis (H_A) allows different partitions to have different topologies; the LHT thus allows for direct testing of topological incongruence in a likelihood framework. Under both H_0 and H_A, other model parameters, e.g., branch lengths and gamma shape parameters, are free to vary among partitions. The LHT compares the likelihood under the null hypothesis (L_0) to the likelihood under the alternative hypothesis (L_A), using the test statistic

$$\delta = \ln L_0 - \ln L_A.$$

We calculate the null distribution of δ by parametric bootstrapping (Huelsenbeck and Bull 1996), although other approaches are possible (Waddell, Kishino, and Ota 2000).

In order to test for topological heterogeneity within the *melanogaster* subgroup, maximum likelihood parameter estimates and likelihood scores were obtained under Topologies I, II, and III for each locus individually, under the GTR+G+I model of substitution, using PAUP*4.0b10 (Swofford 2002), and δ was calculated as above. Parametric bootstrap replicates were generated by simulation under the GTR+G+I model using SeqGen v. 1.1, using the ML parameter estimates for each locus, under the single topology that maximizes the likelihood summed over all loci (Topology II; see Results). *D. pseudoobscura* and *D. ananassae* were not used for this analysis, in order to reduce computational time. *D. eugracilis* and *D. lutescens* are therefore the outgroups for this analysis. Since all inference was conducted on unrooted trees, lack of resolution at this basal node should not be an issue. A similar procedure was used to test for topological heterogeneity between the *melanogaster*, *eugracilis*, and *takahashii* subgroups. Here, *D. pseudoobscura* and *D. ananassae* were used as outgroups, and *D. erecta* was not included. The null distribution was generated using Topology C (see Results).

Tests for Recombination

We tested for intralocus recombination in the common ancestor of the *melanogaster* subgroup, as well as in the common ancestor of *D. eugracilis*, *D. lutescens*, and *D. melanogaster*. To do so, we used a Bayesian Hidden Markov Model (HMM-Bayes) approach (Husmeier and McGuire 2003), as implemented in TOPALi (Milne et al. 2004). Under standard models of DNA evolution, the probability of observing a particular column y_t in a DNA multiple sequence alignment of *n* nucleotides is given by $P(y_t|S, w, \Theta)$, where t is the site label (1 to *n*), S is the tree

topology, w is a vector of branch lengths, and Θ represents the parameters of the chosen model of nucleotide substitution. Whereas it is typically assumed that there is one "true" topology for all *n* sites in a locus, the HMM-Bayes approach allows each site to have a different topology. Topology is treated as a random variable S_t that depends on the site label t. The state space of S_t consists of all possible unrooted topologies for the sequences under consideration, i.e., there are three possible states for any alignment of four sequences. HMM-Bayes uses a Monte Carlo Markov Chain approach to find the state sequence $\hat{S}$ that is best supported by the data. Recombination events are detected as changes in state along the alignment. If recombination has occurred, then different contiguous portions of an alignment may support different tree topologies.

Due to computational limitations, TOPALi only accepts alignments of four sequences. In order to test for intralocus recombination in the common ancestor of the *melanogaster* subgroup, we used gene sequences from *D. melanogaster*, *D. erecta*, *D. yakuba*, and *D. lutescens* as an outgroup. In order to test for intralocus recombination in the common ancestor of *D. eugracilis*, *D. lutescens*, and *D. melanogaster*, we used sequences from these three species, and *D. pseudoobscura* as an outgroup. Alignments for all twelve loci described above were analyzed by HMM-Bayes.

Results

Tests for Saturation and Base Compositional Bias

Using saturation plots, we find no evidence of substitutional saturation for the ingroup taxa at any locus (Figure 6.3 shows two example plots, with distances between ingroups represented by black squares; other data not shown). Thus, excessive homoplasy should not be a major concern for phylogenetic inference within the *D. melanogaster* subgroup. At three loci (*mitch*, *Gld*, and *hb*), there is evidence

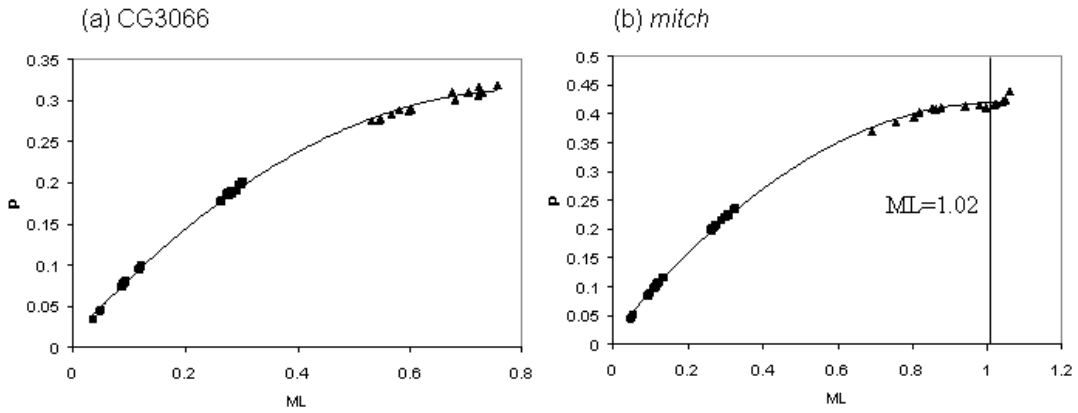


Figure 6.3 Saturation plots of (A) CG3066 and (B) *mitch*. Uncorrected distances (p) between each pair of taxa were plotted against the maximum likelihood corrected distance (ML). Black squares represent distances between ingroup taxa only, while triangles involve at least one outgroup taxon. The fitted line is the best fit second order polynomial, and the vertical line in (B) represents the maximum. To the right of the maximum, substitutional saturation is evident.

for some saturation between the ingroup and outgroup taxa. Base composition equilibrium was rejected at two loci, *ry* ($P < 0.0001$) and *Iris* ($P < 0.0001$). We note that Ko et al. (2003) found little impact of this non-equilibrium base composition on phylogenetic inferences using *ry*; we give further consideration to the potential implications of saturation and non-equilibrium base composition below.

Phylogenetic Inference

Phylogenetic analyses were conducted on all twelve single locus datasets, as well as on a concatenation of all twelve loci. Figure 6.4 summarizes the results of phylogenetic reconstructions for all loci except *Adh*, *Adhr*, *Gld*, and *ry*; results for the latter genes do not differ substantially from those of Ko et al. (2003), and so are not shown here (topologies are described below). Figure 6.5 shows the majority-rule tree and maximum likelihood tree with branch lengths for the concatenated data set. In general, maximum parsimony (MP), maximum likelihood (ML) and Bayesian (B) methods yielded similar tree topologies within a dataset; exceptions are noted below.

Relationships within the melanogaster *subgroup*

Within the *melanogaster* subgroup, phylogenetic reconstructions using single loci yielded several different tree topologies (Figure 6.4). Different reconstruction methods were generally consistent for a given locus. Topology II, according to which *D. erecta* shares a most recent common ancestor with the *D. yakuba- D. teissieri* species pair to the exclusion of *D. melanogaster-D. simulans*, is supported by five of the eight loci presented in Figure 4: *mitch*, CG7415, CG3066, seq211, and *Iris*. With the exception of CG3066, bootstrap scores are high (>80%) for all loci, as are Bayesian clade probabilities (>99%). Topology I, whereby *D. erecta* is basal within the *melanogaster* subgroup, is supported by CG9336. Bootstrap scores and Bayesian

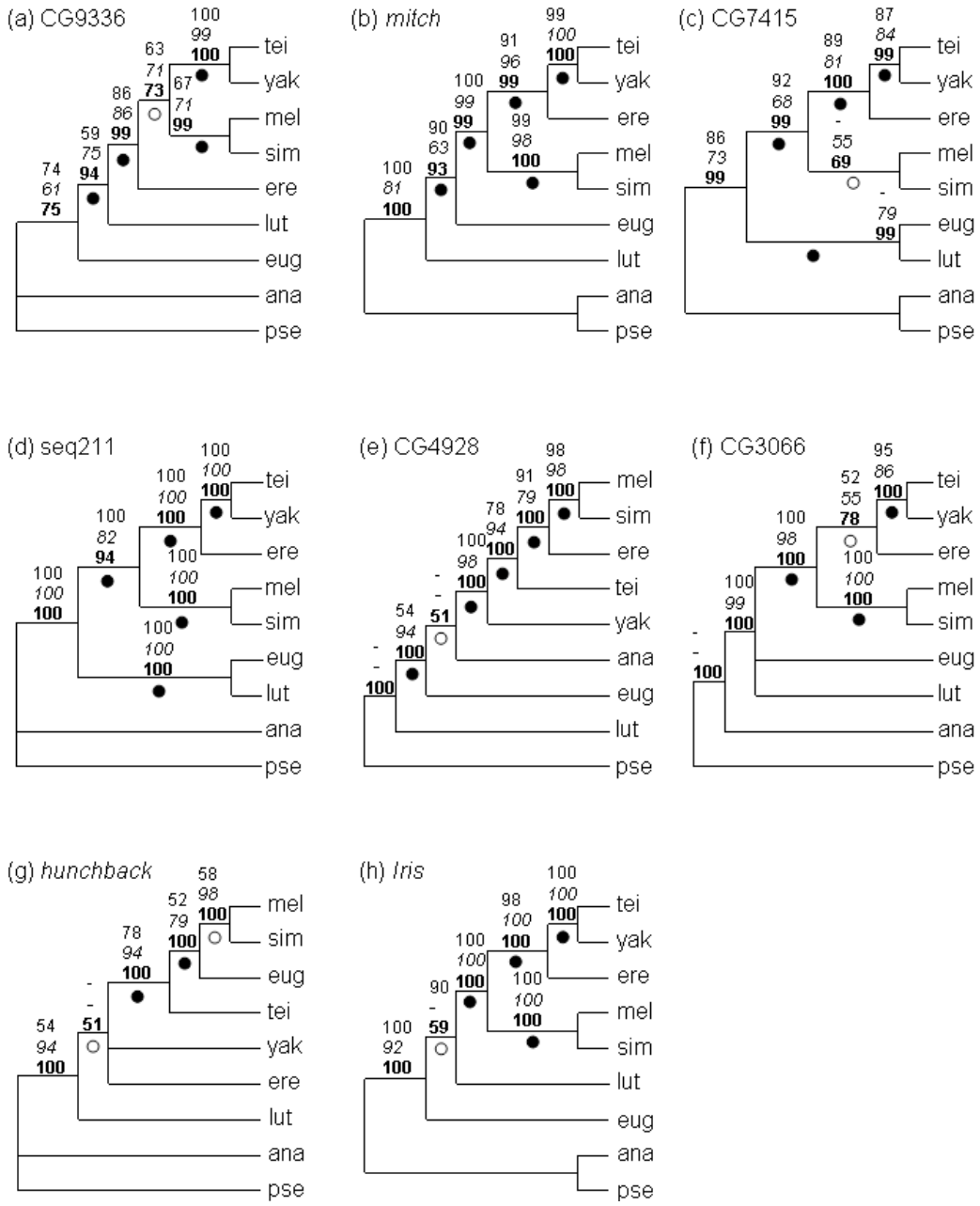


Figure 6.4 Consensus trees for single locus phylogenetic analyses. The numbers above each node indicate, from top to bottom, maximum parsimony bootstrap score (1000 replicates), maximum likelihood bootstrap score (*italic*; 100 replicates), and Bayesian posterior clade probability (**bold**; 500000 generations). For *hunchback*, the three tree construction methods disagree, and the Bayesian consensus tree is shown (see results section). (a) CG9336. (b) *mitch*. (c) CG7415. (d) seq211. (e) CG4928. (f) CG3066. (g) *hunchback*. (h) *Iris*. Zero branch length tests were carried out as described in *Materials and Methods*; open dots represent branches that fail to reject the null hypothesis of zero branch length at a cutoff of 0.05. Black dots represent branches that were tested and do reject the null hypothesis.

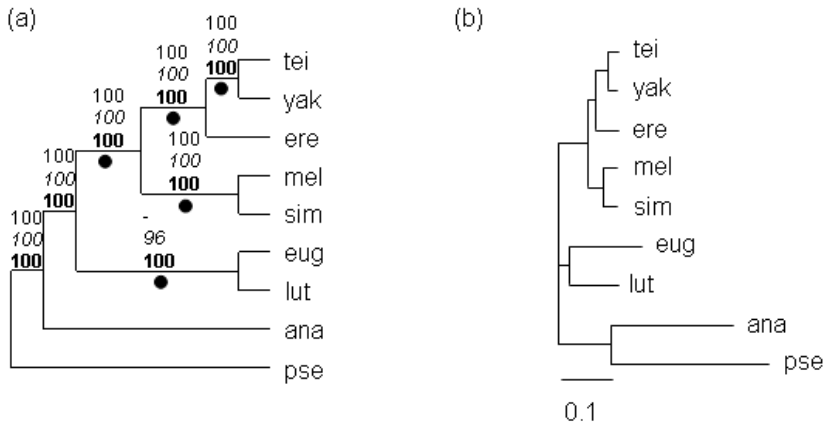


Figure 6.5 (A) Consensus tree for multi-locus analysis. Branch labels are the same as Figure 6.4. (B) Phylogram for the multi-locus analysis. The scale bar represents 0.1 expected substitutions per site.

clade probabilities are, however, relatively low (MP: 63%; ML: 71%; B: 73%). CG4928 supports Topology III, which groups *D. erecta* together with the *D. melanogaster*-*D. simulans* species pair, with fairly strong support (MP: 91%; ML: 79%; B:100%). However, CG4928 also fails to group *D. yakuba* and *D. teissieri* as sister species. Finally, analysis of *hb* fails to support monophyly of the *melanogaster* subgroup, placing *D. eugracilis* as a sister taxon to the *D. melanogaster*-*D. simulans* species pair. Bootstrap scores are quite low for most clades, although Bayesian posterior probabilities are high.

Re-analysis of the four genes studied by Ko et al. (2003) using *D. ananassae* as an additional outgroup did not alter tree topologies within the *melanogaster* subgroup. As in Ko et al. (2003), *Adhr*, *Gld*, and *ry* all support Topology II, whereas *Adh* gives weak support for Topology III (data not shown).

Topology II is strongly supported by a concatenation of all twelve loci examined here (Figure 6.5). Bootstrap scores and the Bayesian clade probability for the (*D. yakuba*/*D. teissieri* + *D. erecta*) grouping are all 100%, indicating robust support for monophyly of this clade.

Relationships between subgroups

Different loci yield different tree topologies with respect to the relationships between *D. lutescens*, *D. eugracilis*, and the *melanogaster* subgroup. Topology A, which places *D. eugracilis* closer to the *melanogaster* subgroup than *D. lutescens*, is supported by two coding loci, *mitch* and CG4928. The degree of support for this branching order varies by method, however, with low maximum likelihood and maximum parsimony bootstrap scores for *mitch* (ML: 63%) and CG4928 (MP: <50%), respectively. All three tree reconstruction methods fail to place *D. ananassae* as an outgroup for CG4928. Topology B, which places *D. lutescens* closer to the

melanogaster subgroup, is weakly supported by CG9336, CG3066, and *Iris*. Maximum parsimony and maximum likelihood bootstrap scores for CG9336 are low (MP: 59%; ML: 75%), while the Bayesian clade probability is high (B: 94%). For CG3066 and *Iris*, bootstrap scores and Bayesian clade probabilities are generally low (CG30666 - MP: 55%; ML: 39%; B: 43%; *Iris* – MP: 90%; ML: 50%; B: 59%). Finally, two loci, CG7415 and seq211, support Topology C, according to which *D. eugracilis* and *D. lutescens* form a group that is monophyletic with respect to the *D. melanogaster* subgroup. This topology is strongly supported by all methods for seq211, but gains mixed support from CG7415.

Ko et al. (2003) found that different tree reconstruction methods yielded incongruent results for *Adh*, *Adhr*, *Gld*, and *ry*. The same general outcome is reached here; different reconstruction methods are consistent only for *ry*, which supports Topology C. No topology is strongly supported by *Adh*. *Adhr* supports Topology A when analyzed using Bayesian analysis, but Topology B under maximum parsimony. Parsimony analysis of *Gld* also supports Topology B, but maximum likelihood and Bayesian analyses support Topology C.

For the concatenated dataset, maximum likelihood and Bayesian methods give strong support to Topology C, with a well supported *D. eugracilis*-*D. lutescens* clade. Maximum parsimony, by contrast, weakly supports Topology B (MP: 66%). We note that inference on the concatenated dataset should be treated with caution, however. For example, one assumption of the Bayesian Markov chain Monte Carlo (MCMC) approach used by MrBayes, that there is a single phylogeny for all sites, is clearly violated in this analysis. Different sites support different tree topologies, and such mixtures of trees are known to confound MCMC methods (Mossel and Vigoda 2005). The behavior of other tree reconstruction methods has not been analyzed for mixture models of this variety, but may be similarly confounded.

Interior branch length tests

Interior branches that fail to reject the null hypothesis of zero branch length at a cutoff of α=0.05 are indicated in Figures 6.4 and 6.5 with an open dot, while branches that were tested but do reject the null hypothesis are marked with a black dot. Within the *melanogaster* subgroup, one or more branches are not significantly different from zero in length for CG9336, CG7415, and *hb*. For most loci, zero branch length is rejected for the branches connecting *D. eugracilis*, *D. lutescens*, and the *melanogaster* subgroup (with the exception of *Iris*).

Tests of incongruence

Applied to all twelve loci considered in this study, the ILD test of Farris et al. (Farris et al. 1995) rejects the null hypothesis of homogeneity ($P < 0.002$). While this result does suggest incongruence among loci, it may be difficult to distinguish rejection due to topological incongruity, rate heterogeneity between loci, or other factors (Dolphin et al. 2000; Barker and Lutzoni 2002; Darlu and Lecointre 2002). Thus, in order to explicitly test for topological incongruence, and to specifically investigate disagreement at the two nodes of interest here, we implemented the LHT of Huelsenbeck and Bull (1996).

Using the LHT, we tested for incongruence with respect to the placement of *D. erecta* in the *melanogaster* subgroup, and the relationships between the *melanogaster*, *eugracilis*, and *takahashii* subgroups (Table 6.2, Figure 6.6), again using all twelve loci. Within the *melanogaster* subgroup, if a single tree is assumed to underlie all loci, Topology II is the maximum likelihood topology (Table 6.2). When the assumption that a single tree underlies all loci is relaxed, such that each locus is allowed any of three possible topologies, an improvement of 7.18 likelihood units is observed (δ =

Table 6.2 Likelihood heterogeneity test - Negative log likelihoods under the GTR+G+I model of substitution. For comparison of Topologies I, II, and III, *D. pseudoobscura* and *D. ananassae* were excluded. For comparison of Topologies A, B, and C, *D. erecta* was excluded. * denotes the maximum likelihood tree. δ_1 is the LHT test statistic for comparison between Topologies I, II, and III, and δ_2 is the LHT test statistic for comparison between Topologies A, B, and C.

	Loci												
Topology	*Adh*	*Adhr*	*ry*	*Gld*	*mitch*	*hb*	Iris	CG3066	CG4928	CG7415	CG9336	seq211	Total
I	1836.41	2394.61	11919.72	4653.70	2407.76	1160.64	7629.75	2967.78	3556.16	2201.19	1041.72*	5996.76	47766.22
II	1836.23	2385.77*	11899.85*	4651.21*	2399.87*	1160.63	7616.29*	2966.24*	3556.16	2196.22*	1041.96	5979.46*	47689.88*
III	1834.20*	2394.61	11921.53	4655.43	2407.73	1160.60*	7631.21	2967.84	3551.27*	2201.32	1041.99	5996.76	47764.51
													δ_1 = 47689.88 – 47682.70 = 7.18 (P = 0.004)
A	2406.89*	3154.87*	15915.77	5881.95	3578.56*	1713.58*	10413.70	4101.74	4618.44*	2883.71	1333.56	8507.18	64509.96
B	2407.79	3156.20	15920.83	5880.39	3580.44	1716.06	10413.23*	4101.40	4620.91	2883.39	1331.48*	8507.18	64519.30
C	2410.39	3156.40	15902.15*	5879.06*	3580.92	1716.05	10414.03	4100.79*	4618.71	2877.92*	1333.56	8492.48*	64482.47*
													δ_2 = 64482.47 - 64469.46 = 13.01 (P < 0.002)

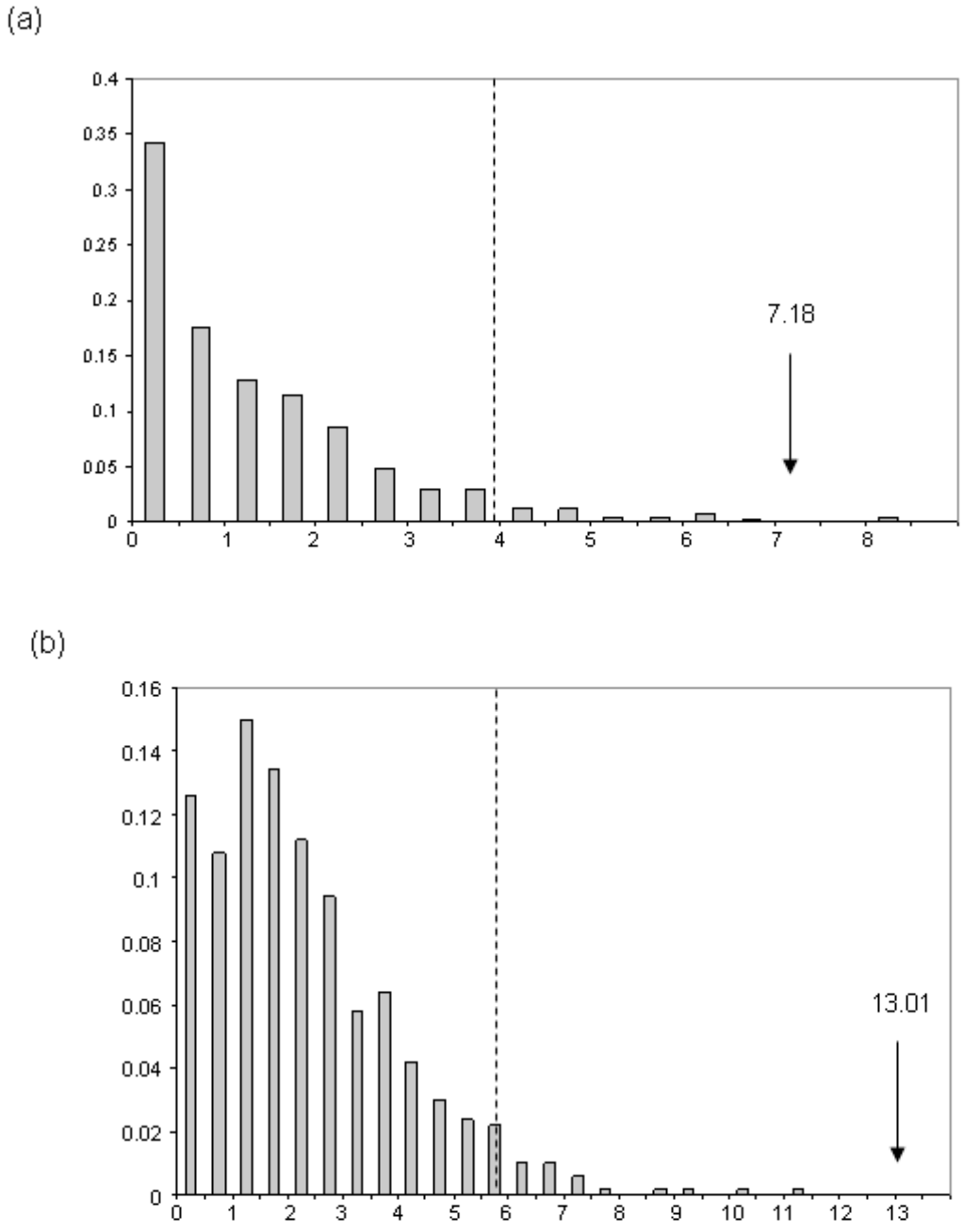


Figure 6.6 Simulated null distributions of δ for tests of topological heterogeneity (A) within the *melanogaster* subgroup and (B) between the *melanogaster*, *eugracilis*, and *takahashii* subgroups. 500 bootstrap replicates were simulated under the hypothesis that a single tree underlies all 12 loci, using maximum likelihood parameter estimates for the original data. The observed values of δ (indicated by a vertical arrow) both fall outside the 95% confidence intervals (dashed line), indicating rejection of the null hypothesis.

7.18; Table 6.2). The null distribution of δ was obtained by parametric bootstrapping on Topology II (Figure 6.6A). Five hundred replicates were performed. Only two replicates had a value of δ more extreme than 7.18 ($P = 0.004$), indicating that the degree of incongruence present in the empirical dataset is unlikely to arise purely from sampling error. In order to identify the source of this incongruence, we excluded single loci from the analysis and re-calculated δ and its null distribution. When CG4928 was excluded, we no longer detected significant incongruence ($\delta = 2.29$; $P = 0.122$), while no other single locus had a similar effect on the test result (data not shown). We suggest that the low rate of substitution at CG4928 (Table 6.1), combined with a short internal branch between *D. erecta* and its relatives, has led to a misleading phylogenetic signal at this locus.

With respect to relationships among subgroups, Topology C provides the best single topology under the null hypothesis (Table 6.2); relaxation of the assumption of a single underlying tree provides an improvement of 13.01 likelihood units ($\delta = 13.01$; Table 6.2). Analysis of 500 simulated datasets suggests that this value of δ is very unlikely to occur by chance ($P < 0.002$; Figure 6.6B). Thus, we reject the null hypothesis that a single topology underlies all twelve loci. Exclusion of single loci did not result in a non-significant test-statistic (data not shown). Moreover, we attempted to assess the impacts of homoplasy, non-equilibrium base composition, and positive selection by excluding loci showing evidence for saturation between outgroup and ingroup taxa, loci rejecting base composition equilibrium, or loci showing evidence for positive selection across numerous taxa (Table 6.3; CG3066 and *Iris*; Swanson et al. 2004; Malik and Henikoff 2005). In each case, the null hypothesis is still rejected, suggesting that none of these potential confounding factors is solely responsible for the observed level of incongruence.

Table 6.3 Values of δ_2 and associated probabilities for subsets of loci.

Subset	Loci removed	δ_2	*P*
All loci	None	13.01	<0.002
Loci with no evidence of saturation between outgroup and ingroup taxa	*Gld*, *hb*, *mitch*	8.17	0.004
Loci with no evidence for base compositional disequilibrium	*ry*, *Iris*	12.21	<0.002
Loci with no evidence for positive selection	CG3066, *Iris*	12.21	<0.002

Evidence for recombination within genes

We used a Bayesian approach to find evidence of recombination events in the common ancestor of the *melanogaster* subgroup, and in the common ancestor of *D. eugracilis*, *D. lutescens*, and *D. melanogaster*. Using TOPALi (Milne et al. 2004), we found statistically significant evidence for recombination at three loci out of twelve tested (Figure 6.7; other data not shown). We find evidence for intralocus recombination in the common ancestor of *D. melanogaster*, *D. eugracilis*, and *D. lutescens* at *mitch* (Fig 6.7A), and at the non-coding locus seq211 (results not shown). In addition, we find evidence for intralocus recombination in the common ancestor of the *melanogaster* subgroup species at *Iris* (results not shown) and at seq211 (Figure 6.7B). We note that this analysis is largely exploratory, since the performance of the HMM-Bayes method has not been rigorously tested under a variety of conditions (including, importantly, situations where homoplasy may arise).

Discussion

Phylogenetic relationships within the melanogaster subgroup

Phylogenetic relationships within the *melanogaster* species group and subgroup have proven difficult to resolve (Kopp and True 2002b; Ko, David, and Akashi 2003; Lewis, Beckenbach, and Mooers 2005; Kopp 2006). In this study, we find strong support for Topology II within the *melanogaster* subgroup, i.e., for the existence of a clade consisting of *D. erecta* and the *D. yakuba*-*D. teissieri* species pair, to the exclusion of *D. melanogaster* and *D. simulans*. In individual locus analyses, eight out of twelve loci support this topology (Figure 6.4; Table 6.2). Moreover, LHT results suggest that one gene, CG4928, is primarily responsible for any statistically significant incongruence between loci; exclusion of CG4928 results in a non-significant test statistic. In addition, analysis of a concatenated dataset consisting of

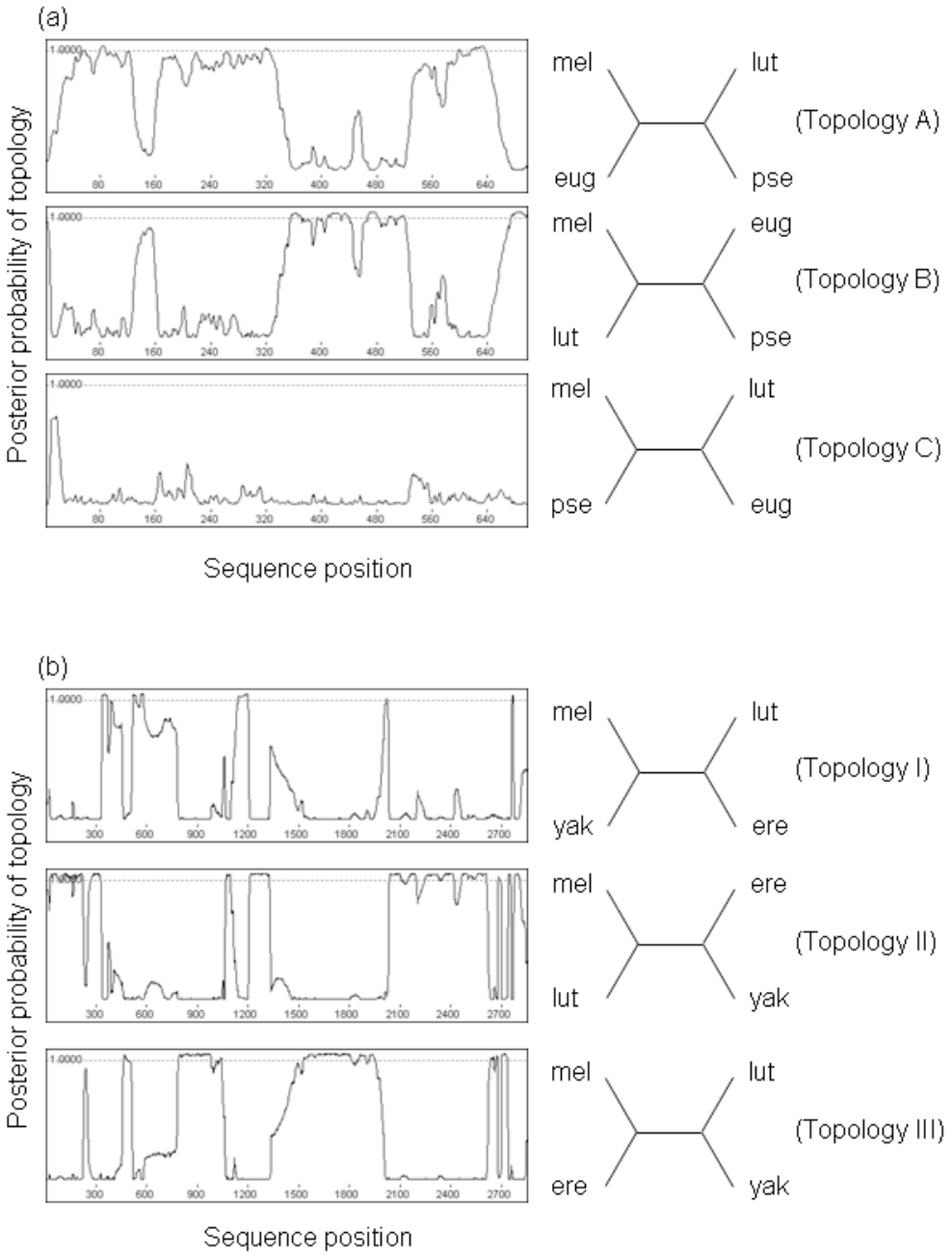


Figure 6.7 Evidence for ancestral lineage sorting with recombination. The plots on the left indicate, across the length of the locus, the posterior probability of each of the topologies shown on the right. (A) *mitch* supports two different tree topologies, A and B, for the relationship between *D. lutescens*, *D. eugracilis*, and the *melanogaster* subgroup. (B) seq211 supports all three possible topologies in the *melanogaster* subgroup.

over 18 kb of sequence provides statistically robust support for Topology II (Figure 6.5). Notably, all multi-locus datasets analyzed to date give the same phylogenetic reconstruction (Kopp and True 2002b; Ko, David, and Akashi 2003), as do numerous independent single locus analyses (Nigro, Solignac, and Sharp 1991; Pelandakis, Higgins, and Solignac 1991; Gailey et al. 2000; Arhontaki et al. 2002). The prevailing alternative hypothesis, whereby *D. erecta* occupies a basal position within the *melanogaster* subgroup (Topology I), is supported by allozyme distance data (Cariou 1987), sequence analysis of *Adh* in early studies (Jeffs, Holmes, and Ashburner 1994; Russo, Takezaki, and Nei 1995) and by biogeographical considerations (Lachaise et al. 1988). The weight of evidence, we argue, is in favor of Topology II.

Phylogenetic relationships between subgroups

The data presented here fail to unambiguously resolve the relationship between *D. eugracilis*, *D. lutescens*, and the *melanogaster* subgroup. In analyses of individual loci and concatenated datasets, tree topology is strongly dependent on choice of locus: of the twelve loci considered in this study, no more than five support any one of the three possible trees (Figure 6.4; Table 6.2). Topology C is strongly supported by the concatenated alignment in model-based analyses, while maximum parsimony yields weak support for Topology B. Similarly, disagreements are common amongst previous studies: Ko et al. (2003), using four loci, argue for Topology A. By contrast, Kopp and True (2002b) find support for Topology B, using data from six loci.

Using the LHT, we find strong evidence for topological incongruence between loci with respect to relationships between subgroups (Figure 6.6; Tables 6.2 and 6.3). This incongruence is not attributable to any single locus. Moreover, we find no evidence that homoplasy, non-equilibrium base composition, or positive selection is

responsible for the signal of incongruence, since tests excluding loci with evidence for any of these factors still reject the null hypothesis (Table 6.3).

Species level polytomies in the melanogaster *species group*

It is well documented that gene trees do not always recapitulate the species tree (e.g., Pamilo and Nei 1988; Wu 1991; Poe and Chubb 2004; Degnan and Salter 2005; Degnan and Rosenberg 2006). One potential reason for such disagreement is sorting of polymorphism in the common ancestor of three or more lineages. Consider the case of three species, A, B, and C, that diverged from a common ancestor, and orthologous gene sequences *a*, *b*, and *c* sampled from these species in the present. Suppose that C diverged first from the common ancestor, and that B subsequently diverged from the lineage leading to A, such that the rooted species tree is appropriately represented as ((A, B), C). In order for the gene tree to recapitulate the species tree, *a* and *b* must find a common ancestor (coalesce) before either coalesces with *c*. The gene tree will fail to accurately represent the species history if *a* coalesces with *c* before either coalesces with *b*, or if *b* coalesces with *c* before either coalesces with *a*.

Pamilo and Nei (1988) showed that, for a neutral locus, the probability *P* that a gene tree has the same topology as the species tree is dependent on only two factors: population size *N*, and time *t* between speciation events. Time to fixation for ancestral polymorphisms is higher for large populations; as such, *P* is smaller for higher values of *N*. A longer period of time between speciation events gives polymorphisms more time to go to fixation; hence, *P* is higher for larger values of *t*. Importantly, then, a short period of time between subsequent speciation events substantially decreases the probability that the gene tree recapitulates the species tree. Towards the limiting case of a polytomy (splitting of an ancestral lineage simultaneously into three or more daughter lineages), the probability that the gene tree has the same topology as the

species tree is only 1/3 in the case of three daughter lineages. Thus, multiple loci sampled from lineages that diverged simultaneously (or nearly so) should show different tree topologies. Incongruence between loci has been cited as evidence for simultaneous or near-simultaneous radiation in, for example, birds (Poe and Chubb 2004) and primates (Ruvolo 1997).

Given this prediction, there are at least two potential species level polytomies in the *melanogaster* species group: One at the root of the *melanogaster* subgroup, and one connecting *D. eugracilis*, *D. lutescens*, and the *melanogaster* subgroup. We can use incongruence between gene trees to test the hypothesis of a species level polytomy, following Ruvolo (1997). Consider three species A, B, and C, with the same *r* independent loci sampled from each one. Suppose that the real species tree is ((A, B), C). For each locus, there are three possible rooted gene trees: ((*a*, *b*), *c*), ((*a*, *c*), *b*), and ((*b*, *c*), *a*). Following Pamilo and Nei (1988), call these topologies α, β, and γ, respectively, and let *i*, *j*, and *k* represent the number of independent loci supporting topologies α, β, and γ. The correct topology is inferred if $i>j$ and $i>k$. We can determine if *i* is greater than the number of loci that would be expected to support topology α under the null hypothesis of a strict polytomy, as follows. Under a polytomy, each topology has an equal probability (1/3) of being realized, such that the probability of obtaining the true topology (α) is 1/3, and the probability of obtaining the wrong topology (β or γ) is 2/3. The probability that *i* or more of the *r* loci support the true topology is therefore given by a sum of binomial probabilities:

$$P(i) = \sum_{n=i}^{r} \binom{r}{n} (1/3)^n (2/3)^{r-n}$$

Failure to reject the null hypothesis indicates that the available data are consistent with polytomy at the species level. Rejection of the null hypothesis, by contrast, suggests that the available data are inconsistent with simultaneous speciation events.

Using gene trees inferred in this and other studies, we can evaluate the probability of a polytomy at the two branch points described above (see Wong et al. 2007 for a list of genes). We note that this approach is approximate, as it fails to take into account uncertainty in individual tree topologies (Satta, Klein, and Takahata 2000). Nonetheless, it should provide some quantitative sense of the robustness of phylogenetic hypotheses. For relationships within the *melanogaster* subgroup, 13 genes support Topology II and 3 do not. Under a polytomy, the probability that 13 or more genes out of 16 will support a single topology is 0.000116; hence, we reject the null hypothesis at this branch point. Our LHT results similarly suggest broad topological congruence between loci concerning relationships within the *melanogaster* subgroup.

By contrast, for relationships between the *melanogaster*, *eugracilis*, and *takahashii* subgroups, 6 genes support Topology A, 5 support Topology C, and 3 support Topology B. The null probability that 6 or more genes out of 14 will support a single topology is 0.31, and hence a species level polytomy cannot be rejected. The data are thus consistent with lineage sorting from the common ancestor of the *melanogaster*, *eugracilis*, and *takahashii* subgroup through closely spaced speciation events. This finding is also consistent with our LHT results, wherein significant incongruence between loci could not be attributed to any single locus or to various potential confounding factors. A recent study (Kopp 2006) argued that the ancient (12-24 mya) divergence of the *melanogaster* species group renders lineage sorting unlikely. However, we note that the relevant time interval for lineage sorting is not the age of divergence, but rather the time *between* closely spaced speciation events. Lineage sorting in the deep history of a clade may still result in incongruence between loci, as subsequent coalescence of alleles within a lineage will not resolve relationships in the ancestral population. We argue that an ancient lineage sorting

event is the best explanation for our results, as well as for Kopp's (2006) finding that relationships between *D. melanogaster*, *D. eugracilis*, and *D. biarmipes* (a close relative of the *takahashii* subgroup) are poorly supported.

We therefore conclude that, within the *melanogaster* subgroup, there is strong support for a monophyletic clade consisting of the *D. yakuba*-*D. teissieri* species pair and the *D. erecta*-*D. orena* species pair (although *D. orena* was not examined in this study, we assume here that it is the sister species to *D. erecta*) . However, we note that the internal branches connecting the *melanogaster*-*simulans*, *teissieri*-*yakuba*, and *erecta*-*orena* species pairs tend to be short (Figure 6.4), and may present some risk of lineage sorting. We argue that Topology C is the best current hypothesis for the speciation history of the *melanogaster*, *eugracilis*, and *takahashii* subgroups, being supported both by partitioned data analysis (Table 6.2) and the combined data (Figure 6.5). Nonetheless, incongruence between loci is widespread, and may be best explained by extensive lineage sorting from a polymorphic ancestor.

Implications for comparative studies

Phylogenetic incongruence within and between loci, of the sort observed in this study, is a potential concern in several lineages of interest to evolutionary biologists. The relationship among humans, chimpanzees, and gorillas is perhaps the best known example. These three primate lineages almost certainly speciated rapidly from a common ancestor, and as a result, different loci provide support for each of three possible rooted tree topologies (Ruvolo 1997; Satta, Klein, and Takahata 2000). Moreover, different sites within a given locus may support different topologies (Satta, Klein, and Takahata 2000). Another well documented example of lineage sorting comes from the *D. simulans* species complex, which includes *D. simulans*, *D. mauritiana*, and *D. sechellia*. Here, speciation is thought to have occurred fairly

recently, such that some ancestral polymorphism is shared between species (Kliman et al. 2000). Only two loci have been identified that support monophyly of alleles within species, and the species relationships that they support are different (Ting, Tsaur, and Wu 2000); (Malik and Henikoff 2005). In addition, full genome sequences are now available for several members of the *melanogaster* subgroup (http://species.flybase.net), and thus will be subject to extensive comparative analyses. We have argued that the lineages giving rise to the sequenced species *D. erecta*, *D. yakuba*, and (*D. melanogaster* + *D. simulans*) may have split in rapid succession, resulting in some lineage sorting and intralocus recombination. Sorting from a polymorphic ancestor, as observed in primates and in several Drosophila lineages, has several implications for comparative studies, three of which we highlight here.

First, phylogenetic inference itself can be complicated by incongruence within and between loci. It is generally acknowledged that single locus analyses are insufficient to resolve species relationships, such that data must be collected from multiple loci in order to make robust inferences. Authors have debated whether multi-locus datasets should be analyzed on a locus-by-locus basis, or whether it is more appropriate and/or powerful to concatenate all loci (e.g., Kluge 1989; Miyamoto and Fitch 1995; Huelsenbeck, Bull, and Cunningham 1996). Advocates of the so-called "total evidence" approach, whereby all data are included in a combined analysis, argue on philosophical grounds about explanatory power (Kluge 1989), or suggest that use of a concatenated dataset allows the dominant phylogenetic signal to "overwhelm" conflicting signals (Rokas et al. 2003). Lineage sorting events may be especially problematic for total evidence approaches, and should be treated with caution generally. For example, a recent study demonstrated that popular MCMC methods perform poorly on datasets containing mixed phylogenetic signals, taking inordinately long to converge on the true tree (Mossel and Vigoda 2005). Moreover, in some cases

where more than three lineages have been affected by lineage sorting, sampling of multiple loci can converge on the wrong species tree in total evidence or locus-by-locus analyses (Degnan and Rosenberg 2006). Such scenarios are especially likely in speciose clades where large population sizes are common (like Drosophila), and necessitate careful analytical procedures. Finally, important information about speciation history can be lost by the use of a concatenated dataset. The presence of extensive incongruence can reveal complex genealogical history, and this will be evident only in multiple single locus analyses and explicit tests of congruence between partitions.

Inference of substitution rates may also be affected by lineage sorting. Consider a case where three species, A, B, and C, arise in rapid succession from a common ancestor, such that polymorphism is shared between them in the early stages of speciation. Here, two mutations in the common ancestor of A, B, and C occurring at partially linked or unlinked sites may give rise to three haplotypes: two haplotypes bearing single mutations, and a recombinant haplotype bearing both. Since polymorphism is initially shared following speciation, a real possibility exists for different haplotypes to go to fixation in each species. Upon sampling gene sequences from A, B, C, we would have to posit recurrent mutation at one of the sites if we were to assume a single tree. Consequently, analyses relying on rate estimates, such as molecular clock inferences and relative rate tests, may be confounded.

Finally, species level polytomies may confound inferences of positive selection, due to the presence of recombination within loci (or between loci for concatenated datasets). Maximum likelihood methods implemented in the popular software package PAML are often used to detect the action of positive selection on coding sequences. These methods are known to be sensitive to recombination; moderate to high levels of recombination can lead to an unacceptably high false

positive rate (Anisimova, Nielsen, and Yang 2003). The increased false positive rate associated with recombination may result from the assumption that the rate of synonymous substitution is homogeneous across all sites (nonsynonymous substitution rates are allowed to vary between codons), or from the use of an incorrect tree for some sites (Anisimova, Nielsen, and Yang 2003). Although lineage sorting in a deep ancestor has not been explicitly investigated as a source of error in PAML and related analyses, it may have confounding effects.

We suggest several approaches to circumvent inferential problems stemming from ancestral lineage sorting and recombination. First, where possible, we recommend care in the choice of taxa used for studies of molecular evolution. Where three lineages are suspected to have arisen in quick succession from their common ancestor, no more than two should be chosen for analyses dependent on accurate estimates of the substitution rate. In this way, the possibility of all four possible arrangements (including outgroup species) of two biallelic sites appearing in the sample due to recombination is eliminated. Polytomies involving more than three lineages should be treated with extra caution.

Moreover, given that ancestral recombination can lead to conflicting phylogenetic signals and inflation of rate estimates *within* a locus (Satta, Klein, and Takahata 2000; this study), analytical methods that explicitly account for recombination (e.g., Wilson and McVean 2006) should be used where such histories are a concern. Alternatively, datasets should be examined for intragenic recombination, especially for lineages with histories known to be problematic. Inference may then be conducted on segments supporting the same topology.

REFERENCES

Anisimova, M., R. Nielsen, and Z. Yang. 2003. Effect of recombination on the accuracy of the likelihood method for detecting positive selection at amino acid sites. Genetics **164**:1229-1236.

Arhontaki, K., E. Eliopoulos, G. Goulielmos, P. Kastanis, S. Tsacas, M. Loukas, and F. Ayala. 2002. Functional constraints of the Cu,Zn superoxide dismutase in species of the *Drosophila melanogaster* subgroup and phylogenetic analysis. J Mol Evol **55**:745-756.

Barker, F. K., and F. M. Lutzoni. 2002. The utility of the incongruence length difference test. Syst Biol **51**:625-637.

Bray, N., and L. Pachter. 2004. MAVID: constrained ancestral alignment of multiple sequences. Genome Res **14**:693-699.

Cariou, M. L. 1987. Biochemical phylogeny of the eight species in the *Drosophila melanogaster* subgroup, including *D. sechellia* and *D. orena*. Genet Res **50**:181-185.

Darlu, P., and G. Lecointre. 2002. When does the incongruence length difference test fail? Mol Biol Evol **19**:432-437.

Degnan, J. H., and N. A. Rosenberg. 2006. Discordance of species trees with their most likely gene trees. PLoS Genet **2**:e68.

Degnan, J. H., and L. A. Salter. 2005. Gene tree distributions under the coalescent process. Evolution Int J Org Evolution **59**:24-37.

Dolphin, K., R. Belshaw, C. D. Orme, and D. L. Quicke. 2000. Noise and incongruence: interpreting results of the incongruence length difference test. Mol Phylogenet Evol **17**:401-406.

Farris, J. S., M. Kallersjo, A. G. Kluge, and C. Bult. 1995. Testing significance of incongruence. Cladistics **10**:315-319.

Felsenstein, J. 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. J Mol Evol **17**:368-376.

Gailey, D. A., S. K. Ho, S. Ohshima, J. H. Liu, M. Eyassu, M. A. Washington, D. Yamamoto, and T. Davis. 2000. A phylogeny of the Drosophilidae using the sex-behaviour gene fruitless. Hereditas **133**:81-83.

Goldman, N., and S. Whelan. 2000. Statistical tests of gamma-distributed rate heterogeneity in models of sequence evolution in phylogenetics. Mol Biol Evol **17**:975-978.

Huelsenbeck, J. P., and J. J. Bull. 1996. A likelihood ratio test to detect conflicting phylogenetic signal. Syst Biol **45**:92-98.

Huelsenbeck, J. P., J. J. Bull, and C. W. Cunningham. 1996. Combining data in phylogenetic analysis. Trends Ecol Evol **11**:152-158.

Huelsenbeck, J. P., and F. Ronquist. 2001. MRBAYES: Bayesian inference of phylogenetic trees. Bioinformatics **17**:754-755.

Husmeier, D., and G. McGuire. 2003. Detecting recombination in 4-taxa DNA sequence alignments with Bayesian hidden Markov models and Markov chain Monte Carlo. Mol Biol Evol **20**:315-337.

Jeffs, P. S., E. C. Holmes, and M. Ashburner. 1994. The molecular evolution of the alcohol dehydrogenase and alcohol dehydrogenase-related genes in the *Drosophila melanogaster* species subgroup. Mol Biol Evol **11**:287-304.

Kastanis, P., E. Eliopoulos, G. N. Goulielmos, S. Tsakas, and M. Loukas. 2003. Macroevolutionary relationships of species of *Drosophila melanogaster* group based on mtDNA sequences. Mol Phylogenet Evol **28**:518-528.

Kliman, R. M., P. Andolfatto, J. A. Coyne, F. Depaulis, M. Kreitman, A. J. Berry, J. McCarter, J. Wakeley, and J. Hey. 2000. The population genetics of the origin

and divergence of the *Drosophila simulans* complex species. Genetics **156**:1913-1931.

Kluge, A. G. 1989. A concern for evidence and a phylogenetic hypothesis of relationships among Epicrates (Boidae, Serpentes). Syst Zool **38**:7-25.

Ko, W. Y., R. M. David, and H. Akashi. 2003. Molecular phylogeny of the *Drosophila melanogaster* species subgroup. J Mol Evol **57**:562-573.

Kopp, A. 2006. Basal relationships in the *Drosophila melanogaster* species group. Mol Phylogenet Evol **39**:787-798.

Kopp, A., and J. R. True. 2002a. Evolution of male sexual characters in the oriental *Drosophila melanogaster* species group. Evol Dev **4**:278-291.

Kopp, A., and J. R. True. 2002b. Phylogeny of the Oriental *Drosophila melanogaster* species group: a multilocus reconstruction. Syst Biol **51**:786-805.

Lachaise, D., M.-L. Cariou, J. R. David, F. Lemeunier, and M. Ashburner. 1988. Historical biogeography of the *D. melanogaster* species subgroup. **Evol Biol**:159-226.

Lemeunier, F., J. R. David, and L. Tsacas. 1986. The *melanogaster* species group. Pp. 148-256 *in* M. Ashburner, H. L. Carson, and J. N. Thompson, eds. *Genetics and Biology of Drosophila*. Academic Press, New York.

Lewis, R. L., A. T. Beckenbach, and A. O. Mooers. 2005. The phylogeny of the subgroups within the melanogaster species group: likelihood tests on COI and COII sequences and a Bayesian estimate of phylogeny. Mol Phylogenet Evol **37**:15-24.

Malik, H. S., and S. Henikoff. 2005. Positive selection of Iris, a retroviral envelope-derived host gene in *Drosophila melanogaster*. PLoS Genet **1**:e44.

McDonald, J. H., and M. Kreitman. 1991. Adaptive protein evolution at the Adh locus in Drosophila. Nature **351**:652-654.

Milne, I., F. Wright, G. Rowe, D. F. Marshall, D. Husmeier, and G. McGuire. 2004. TOPALi: software for automatic identification of recombinant sequences within DNA multiple alignments. Bioinformatics **20**:1806-1807.

Miyamoto, M. M., and W. M. Fitch. 1995. Testing species phylogenies and phylogenetic methods with congruence. Syst Biol **44**:64-76.

Mossel, E., and E. Vigoda. 2005. Phylogenetic MCMC algorithms are misleading on mixtures of trees. Science **309**:2207-2209.

Nigro, L., M. Solignac, and P. M. Sharp. 1991. Mitochondrial DNA sequence divergence in the *melanogaster* and oriental species subgroups of Drosophila. J Mol Evol **33**:156-162.

Pamilo, P., and M. Nei. 1988. Relationships between gene trees and species trees. Mol Biol Evol **5**:568-583.

Pelandakis, M., D. G. Higgins, and M. Solignac. 1991. Molecular phylogeny of the subgenus Sophophora of Drosophila derived from large subunit of ribosomal RNA sequences. Genetica **84**:87-94.

Poe, S., and A. L. Chubb. 2004. Birds in a bush: five genes indicate explosive evolution of avian orders. Evolution Int J Org Evolution **58**:404-415.

Prud'homme, B., N. Gompel, A. Rokas, V. A. Kassner, T. M. Williams, S. D. Yeh, J. R. True, and S. B. Carroll. 2006. Repeated morphological evolution through cis-regulatory changes in a pleiotropic gene. Nature **440**:1050-1053.

Richards, S., Y. Liu, B. R. Bettencourt, P. Hradecky, S. Letovsky, R. Nielsen, K. Thornton, M. J. Hubisz, R. Chen, R. P. Meisel, O. Couronne, S. Hua, M. A. Smith, P. Zhang, J. Liu, H. J. Bussemaker, M. F. van Batenburg, S. L. Howells, S. E. Scherer, E. Sodergren, B. B. Matthews, M. A. Crosby, A. J. Schroeder, D. Ortiz-Barrientos, C. M. Rives, M. L. Metzker, D. M. Muzny, G. Scott, D. Steffen, D. A. Wheeler, K. C. Worley, P. Havlak, K. J. Durbin, A.

Egan, R. Gill, J. Hume, M. B. Morgan, G. Miner, C. Hamilton, Y. Huang, L. Waldron, D. Verduzco, K. P. Clerc-Blankenburg, I. Dubchak, M. A. Noor, W. Anderson, K. P. White, A. G. Clark, S. W. Schaeffer, W. Gelbart, G. M. Weinstock, and R. A. Gibbs. 2005. Comparative genome sequencing of Drosophila pseudoobscura: chromosomal, gene, and cis-element evolution. Genome Res **15**:1-18.

Rokas, A., B. L. Williams, N. King, and S. B. Carroll. 2003. Genome-scale approaches to resolving incongruence in molecular phylogenies. Nature **425**:798-804.

Ronquist, F., and J. P. Huelsenbeck. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics **19**:1572-1574.

Russo, C. A., N. Takezaki, and M. Nei. 1995. Molecular phylogeny and divergence times of drosophilid species. Mol Biol Evol **12**:391-404.

Ruvolo, M. 1997. Molecular phylogeny of the hominoids: inferences from multiple independent DNA sequence data sets. Mol Biol Evol **14**:248-265.

Satta, Y., J. Klein, and N. Takahata. 2000. DNA archives and our nearest relative: the trichotomy problem revisited. Mol Phylogenet Evol **14**:259-275.

Schawaroch, V. 2002. Phylogeny of a paradigm lineage: the *Drosophila melanogaster* species group (Diptera: Drosophilidae). Biol J Linn Soc Lond **76**:21-37.

Schlotterer, C., M. T. Hauser, A. von Haeseler, and D. Tautz. 1994. Comparative evolutionary analysis of rDNA ITS regions in Drosophila. Mol Biol Evol **11**:513-522.

Slowinski, J. B. 2001. Molecular polytomies. Mol Phylogenet Evol **19**:114-120.

Swanson, W. J., A. Wong, M. F. Wolfner, and C. F. Aquadro. 2004. Evolutionary expressed sequence tag analysis of Drosophila female reproductive tracts identifies genes subjected to positive selection. Genetics **168**:1457-1465.

Swofford, D. 2002. PAUP*. Phylogenetic analysis using parsimony (*and other methods). Version 4. Sinauer Associates, Sunderland, Mass.

Ting, C. T., S. C. Tsaur, and C. I. Wu. 2000. The phylogeny of closely related species as revealed by the genealogy of a speciation gene, Odysseus. Proc Natl Acad Sci U S A **97**:5313-5316.

Waddell, P. J., H. Kishino, and R. Ota. 2000. Rapid evaluation of the phylogenetic congruence of sequence data using likelihood ratio tests. Mol Biol Evol **17**:1988-1992.

Wilson, D. J., and G. McVean. 2006. Estimating diversifying selection and functional constraint in the presence of recombination. Genetics **172**:1411-1425.

Wittkopp, P. J., J. R. True, and S. B. Carroll. 2002. Reciprocal functions of the Drosophila yellow and ebony proteins in the development and evolution of pigment patterns. Development **129**:1849-1858.

Wong, A., A. D. Jensen, J. E. Pool, and C. F. Aquadro. 2007. Phylogenetic incongruence in the *Drosophila melanogaster* species group. Mol Phy Evol 43: 1138-1150.

Wu, C. I. 1991. Inferences of species phylogeny in relation to segregation of ancient polymorphisms. Genetics **127**:429-435.

Yang, Y., Y. P. Zhang, Y. H. Qian, and Q. T. Zeng. 2004. Phylogenetic relationships of *Drosophila melanogaster* species group deduced from spacer regions of histone gene H2A-H2B. Mol Phylogenet Evol **30**:336-343.

Yang, Z. 1994. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. J Mol Evol **39**:306-314.

Yang, Z., R. Nielsen, N. Goldman, and A. M. Pedersen. 2000. Codon-substitution models for heterogeneous selection pressure at amino acid sites. Genetics **155**:431-449.

APPENDIX

A ROLE FOR THE *DROSOPHILA MELANOGASTER* SEMINAL FLUID LECTIN ACP29AB IN FEMALE SPERM STORAGE[1]

Abstract

Females of many animal species store sperm for periods of a few hours to years. Female sperm storage has important functional and evolutionary consequences, yet relatively little is known of the molecular basis of this phenomenon. In this study, we report that the *Drosophila melanogaster* seminal fluid protein Acp29AB is required for the normal maintenance of sperm in storage. Consistent with this role, Acp29AB localizes to the female sperm storage organs following mating, although it does not appear to associate tightly with sperm. Acp29AB is a predicted lectin, suggesting that sugar-protein interactions are important for *D. melanogaster* sperm storage, much as they are in many mammals. Previous association studies have found an effect of *Acp29AB* genotype on a male's sperm competitive ability; our findings suggest that differences in sperm storage may underlie differences in sperm competition.

Introduction

The acts of insemination and fertilization are temporally separate events in many animal species. Rather than traveling immediately to the waiting ovum, sperm are typically held in storage, often in specialized regions of the female reproductive tract. In most mammals, for example, sperm are stored in an oviductal reservoir for a period of a few hours or days (reviewed in Suarez 2002; Rodriguez-Martinez 2007). Moreover, many insects store sperm in highly specialized storage organs, with sperm surviving for weeks (as in Drosophila; e.g. Bloch Qazi, Heifetz, and Wolfner 2003) to many years (as in some social hymenopterans).

Sperm storage has a number of important functional and evolutionary consequences. From a functional perspective, storage of sperm is often a vital component of reproduction: Studies in mammals suggest that sperm storage in the oviductal reservoir helps to prevent polyspermy (see Suarez 2002), and that it may facilitate control over the process of sperm activation (Suarez 2002; Rodriguez-Martinez 2007). In insects, female sperm storage may reduce the number of potentially costly matings required for full female fecundity, and allows the fertilization of hundreds or thousands of eggs from one or a few matings (Bloch Qazi, Heifetz, and Wolfner 2003). In *Drosophila melanogaster*, for example, females store ~700-1000 of the 4000 sperm received in a single mating, and use approximately 400 for fertilization over a period of about two weeks.

In addition to being important for successful reproduction, the phenomenon of sperm storage can have profound evolutionary consequences. In combination with multiple mating by females (polyandry), sperm storage can generate strong selective pressures on males. If sperm from different males are simultaneously present in the reproductive tract of a single female, then any trait that grants greater fertilization success to one male over his competitor(s) will be favored by selection. Multiple mating and sperm storage thus create the potential for at least two types of selective regime: Sperm competition, whereby sperm from different males present in the same female at the same time compete over ova (Birkhead and Møller 1998; Parker 1998; Simmons 2001), and cryptic female choice, whereby a female preferentially uses sperm from one male over another (Eberhard 1996). Consequently, sperm competition and cryptic female choice are thought to underlie such diverse phenomena as sperm gigantism, sperm polymorphism, and the rapid evolution of some reproductive proteins (Swanson and Vacquier 2002).

[1] This work was a collaboration between myself, Shannon Albright, Ravi Ram Kristipati, Shuqing Ji, Jon Giebel, Anthony Fiumera, and Mariana Wolfner.

While the physiological mechanisms of sperm storage have been well-described in several systems (Suarez 2002; Bloch Qazi, Heifetz, and Wolfner 2003; Adams and Wolfner 2007; Rodriguez-Martinez 2007), and its evolutionary implications explored in detail, the identities of the molecules responsible for sperm storage are still relatively mysterious. Work in mammals and in Drosophila has, however, begun to identify both male and female molecular contributions to sperm storage.

In *Drosophila melanogaster*, females store sperm in two types of organ: the long, coiled seminal receptacle, and the paired spermathecae. It is thought that sperm from the seminal receptacle are used first, with the spermathecae acting as long-term storage organs. Interestingly, the spermathecae appear to secrete substances required for sperm survival in both types of storage organ, since sperm stored in the seminal receptacles of *lozenge* mutant females (which lack spermathecae) have reduced viability. The identities of such spermathecal factors have not been determined. It is known, however, that the enzyme Glucose dehydrogenase (Gld) is required for normal sperm storage, since *Gld* mutant females store suffer defects in sperm storage and in the release of sperm from storage (Iida and Cavener 2004). Recent studies have identified a number of genes expressed in the sperm storage organs (Allen and Spradling 2008), which should lead to further progress in identifying female-expressed genes involved in sperm storage.

A number of male-expressed genes have known roles in sperm storage in *D. melanogaster*. The carboxylesterase Est-6, which is produced in the male ejaculatory duct and bulb, appears to be involved the release of sperm from storage (Gilbert and Richmond 1981). Moreover, mutational, RNAi, and directed cell-ablation studies have shown that seminal fluid proteins produced in the male accessory gland (Acps, for Accessory gland proteins) are necessary for the entry of sperm into storage, as well as for their maintenance and release from storage (Tram and Wolfner 1998; Neubaum and Wolfner 1999; Bloch Qazi and Wolfner 2003; Ram and Wolfner 2007). Several specific Acps have been identified that play important roles in sperm storage: Acp36DE, a large glycoprotein, is required for sperm entry into storage (Bloch Qazi and Wolfner 2003), and plays a role in sperm competition (Clark et al. 1995; Chapman et al. 2000). In addition, an additional four Acps – the lectins CG1652 and CG1656, the cysteine rich secretory protein (CRISP) CG17575, and the protease homolog CG9997 – were recently shown to be necessary for the release of sperm from storage (Ram and Wolfner 2007).

The finding that predicted lectins (a class of sugar-binding proteins) are involved in sperm storage in *D. melanogaster* raises interesting parallels to sperm storage in other animals. In a number of mammals, sperm are stored for several hours in an oviductal reservoir, consisting of sperm bound tightly to the epithelium (e.g., (Suarez and Osman 1987). Carbohydrates mediate the adherence of sperm to the epithelium, with different sugars playing important roles in different species (e.g., DeMott, Lefebvre, and Suarez 1995; Lefebvre et al. 1995; Lefebvre, Lo, and Suarez 1997; Ekhlasi-Hundrieser et al. 2005). In cows, for example, biochemical studies suggest that fucose residues conjugated to annexins act as oviductal receptors for sperm, with several sperm-bound seminal proteins recognizing the fucose moiety (Gwathmey et al. 2006; Ignotz, Cho, and Suarez 2007).

In this study, we provide evidence that the seminal fluid protein Acp29AB, another predicted lectin, contributes to sperm storage in *D. melanogaster*. The *Acp29AB* gene was first identified in an accessory gland cDNA library screen (Wolfner et al. 19997), and is predicted to encode a secreted Ca^{2+}-dependent (C-type) lectin (Wolfner et al. 1997; Mueller et al. 2004). Three lines of evidence suggest a role for Acp29AB in sperm storage. First, Clark et al. (1995) and Fiumera et al. (2005) found associations between naturally occurring alleles at the *Acp29AB* locus and a male's sperm competitive ability, a pattern that could be generated by differences in sperm storage between males bearing different *Acp29AB* alleles. Second, consistent with a role for Acp29AB in sperm competition and/or cryptic female choice, Aguadé (1999) found evidence for positive selection on *Acp29AB*, with an excess of amino acid substitutions between *D.*

melanogaster and its close relative *D. simulans*. Finally, Acp29AB's predicted molecular function as a lectin suggests a possible role in sperm storage, given the role of protein-sugar interactions in sperm storage in mammals (see above), and in sperm-egg interactions in many animals (for a review see Mengerink and Vacquier 2001).

We show that Acp29AB localizes to the female sperm storage organs following mating, consistent with a role for this protein in sperm storage. Moreover, sperm from males bearing the apparent loss of function mutation *Acp29AB*^{1} are not maintained efficiently in storage. *Acp29AB*^{1} mutant males also perform poorly in sperm competition, possibly as a consequence of reduced numbers of stored sperm.

Materials and Methods

Fly handling and rearing

All fly lines were maintained on yeast-glucose media at room temperature on 12 hour light:12 hour dark cycles. Males and virgin females were aged for 3-5 days before mating and/or dissection.

Production and affinity purification of anti-Acp29AB antibodies

A N-terminal His-tagged fusion of amino acids 22-128 (of 234) of Acp29AB was produced in *E. coli* using the vector pBAD-DEST49 (Invitrogen), according to standard protocols (Ravi Ram, Ji, and Wolfner 2005). Following SDS-PAGE, the 29kD fusion protein was gel purified and used to inject rabbits; rabbit injection and boosts were carried out by CRAR/Cornell. Anti-Acp29AB N-terminal antibodies from the rabbit serum were affinity purified against the His Patch-thioredoxin-N terminal Acp29AB fusion protein using a strip purification protocol as described in (Monsma, Harada, and Wolfner 1990).

Acp29AB transfer and localization

In order to confirm transfer of Acp29AB to females, and to localize it in female reproductive tracts, whole female reproductive tracts or portions thereof were dissected in Ringer's solution. Whole reproductive tracts were homogenized in Ringer's solution with protease inhibitors (Roche), and SDS sample buffer was added. Sperm storage organs were homogenized in SDS sample buffer (Park and Wolfner 1995; Ravi Ram, Ji, and Wolfner 2005). Hemolymph sample collection and Western blotting were performed according to Ravi Ram and Wolfner 2005.

Identification of an Acp29AB mutant

In order to identify an Acp29AB mutant, we screened ~4500 lines from the Zuker collection (Koundakjian et al. 2004) bearing EMS induced mutations on a *cn bw* second chromosome for altered amounts of Acp29AB protein. Total protein was extracted from males homozygous for EMS-mutagenized chromosomes. For each line, two whole 3-5 day old (where possible) mutant male or control female (negative control) adult flies were ground in TE (50mM Tris-HCl and 10mM EDTA, pH 7.5), and then SDS sample buffer was added to 2x concentration. Following SDS-PAGE and Western transfer, proteins were cross-linked to the membrane with a 1XPBS 0.5% glutaraldehyde solution. Western blotting was performed according to standard protocols using α-Acp29AB antibodies at 1:250 concentration.

Since the Acp29AB mutant line identified in this manner (*Acp29AB*^{1}) carries a linked spermatogenesis mutation (data not shown), all experiments described below were carried out using *Acp29AB*^{1} hemizygotes over the deficiency Df(2L)ED611 (Ryder et al. 2007), unless mentioned otherwise. Sibling *Acp29AB*^{1}/CyO males were used as controls.

Nucleic acid and sequencing analysis

For Northern analysis, total RNA was isolated from adult male *Drosophila* using Trizol reagent (Gibco BRL) and then poly $(A)^{+}$ purified using the PolyATtract mRNA

isolation kit (Promega). Northern blots were prepared using standard procedures as described (Current Protocols REF). ~10 μg of RNA was run per lane and the blot was probed with random-primed Acp29AB or, as a control, β1-tubulin (Bialojan, Falkenburg, and Renkawitz-Pohl 1984) DNA probes. For reverse transcriptase PCR, cDNA was prepared from RNA extracts of 30 each of homozygous *Acp29AB*^{1}, *Acp29AB*^{1}/CyO, and CS male flies and 30 CS females following the Superscript (Gibco) instructions. Primers that amplify full length *Acp29AB* and *Acp76A* were used in PCR reactions against previously-mentioned cDNA samples and a CS male genomic control. For sequence analysis of candidate mutants, the *Acp29AB* coding region was PCR-amplified using primers 5Acp29AB (5' GGATCTCACACGCTTGAAATCTTCC 3') and 3Acp29AB (5' GTGGGTGTTGCAAATAGCTTGAATGA 3') from genomic DNA. Both strands of the amplified products were sequenced directly by Cornell's BRC with the following primers: 1875Acp29AB (5' CAAATCTGGCCACAAATATACATAACC 3'), 1083Acp29AB (5' GCCAACTTTCTCGAATCGTCTCAT 3'), and 1107Acp29AB (5' GAGACGATTCGAGAAAGTTGGCT 3').

Phenotypic analysis of the Acp29AB^{1} *mutant*

For analysis of the effects of the presence or absence of Acp29AB on female remating behavior, 3-5 day old virgin CS females were allowed to mate with *Acp29AB*^{1}/ Df(2L)ED611 males or control siblings for 1 hour, after which males were discarded. Females that mated successfully during this time were then given access to a new CS male 1 or 4 days later, again for 1 hour, during which time matings were observed and counted. Egg-laying, fertility, and hatchability were assayed as described in Ravi Ram and Wolfner (2007) for a period of 10 days post-mating.

The role of Acp29AB in sperm competition was assayed by the estimation of two relevant parameters: P1, the proportion of offspring sired by a male when he is the first of two males to mate, and P2, the proportion of offspring sired by a male when he is the second of two males to mate. For the estimation of P1, 3-5 day old *cn bw* females (which have white eyes) were first mated to *Acp29AB*^{1} males or their control siblings; these males were heterozygous for both *cn* and *bw*, and hence produced half white-eyed (*cn bw Acp29AB*^{1} / *cn bw*) and half red-eyed (*cn bw* / Df(2L)ED611 or *cn bw* / CyO) progeny. Two days later, the same females were allowed to mate with *cn bw* males, whose progeny from this mating were all white-eyed; matings were observed, and only females that successfully mated both times were kept. Progeny from eggs laid ten days after the second mating were scored for eye color. P1 was estimated as twice the number of red-eyed progeny (since half of the first male's progeny had white eyes) divided by total progeny. Estimation of P2 was conducted in a similar manner, except that females were first mated to a *cn bw* male, and subsequently to *Acp29AB*^{1} males or their control siblings.

Statistical Analyses

Statistical analyses were performed using R version 2.5.1 (Team 2008) or JMP version 5.1.

Results

Transfer and localization of Acp29AB

Western blot analysis using affinity purified antibodies against Acp29AB detects a 29kD protein in extracts of CS male accessory glands, but not in extracts of DTA-E males that lack accessory gland main cells (Kalb, DiBenedetto, and Wolfner 1993) (Figure 1, Panel A), consistent with the initial identification of *Acp29AB* as a male accessory gland specific transcript (Wolfner et al. 1997). The predicted molecular weight of Acp29AB (excluding the putative signal peptide) is 24.8kD. The higher apparent molecular weight may be due to post-translational modifications, such as glycosylation; other Acps are known to be glycosylated (Monsma, Harada, and Wolfner 1990; Bertram,

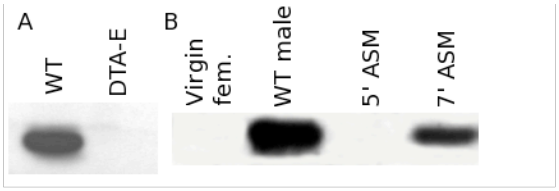


Figure 1. Acp29AB is produced in the male accessory glands (A), and is transferred to females during mating (B). Panel A: Western blots using α-Acp29AB antibodies detect Acp29AB in wild-type (CS) male accessory glands, but not in genital tracts from DTA-E males, which lack the main secretory cells of the accessory gland. Panel B: Reproductive tracts of virgin females and females 5 minutes ASM contain no detectable Acp29AB, but Acp29AB is detected in reproductive tracts from 10 females 7 minutes ASM.

Neubaum, and Wolfner 1996), and sequence data predicts that Acp29AB can be N-glycosylated at amino acids 61 and 164 (Wolfner et al. 1997). Following mating, Acp29AB is transferred from the male to the female: Acp29AB is absent from the reproductive tracts of virgin females and females 5 minutes after the start of mating (ASM), but is transferred to the female reproductive tract by 7 minutes ASM (Figure 1, Panel B).

To examine the targets of Acp29AB in the mated female, we performed Western blot analyses of Acp29AB in extracts of mated female spermathecae, the sperm mass, and mated female hemolymph. We find that Acp29AB does localize to the spermathecae 1 hour ASM (Figure 2, panel A); similar results were obtained 45 minutes ASM (data not shown). Given the localization of Acp29AB to the spermathecae, we performed sperm binding assays to determine if Acp29AB is tightly bound to sperm (Figure 2, panel B). We found no evidence that Acp29AB associates tightly with sperm, since it was never observed in the pelleted sperm fraction. However, Acp29AB is present in the sperm mass, the mass of sperm and seminal fluid transferred to the female during copulation (Figure 2, panel C). Finally, we found that a small quantity of Acp29AB is present in the hemolymph of mated females (Figure 2, panel D); several Acps have been shown to enter the mated female's hemolymph, which would allow them to elicit their effects via the neuroendocrine system (Monsma, Harada, and Wolfner 1990; Lung and Wolfner 1999).

Identification of an Acp29AB mutant

We performed Western blot analysis of protein extracts of whole males from each of ~4500 second chromosome EMS-mutagenized fly lines (Koundakjian et al. 2004); kindly provided by Dr. Charles Zuker) in order to identify lines whose males either lacked Acp29AB or made an altered version of the protein. From initial Western blots, we identified 13 potential *Acp29AB* mutant lines (data not shown). Upon retesting, one line, 83-65, consistently showed no detectable Acp29AB protein; we designate this mutant allele *Acp29AB*^{1} (Figure 3, panel A).

Sequence analysis of the open reading frame of *Acp29AB*^{1} shows that there is a single base pair deletion (A602) that disrupts the reading frame within *Acp29AB*^{1}'s predicted carbohydrate recognition domain (CRD) (Figure 3, panel B). Due to this frameshift mutation, all amino acids after 203 are misencoded, and the polypeptide chain is predicted to be 242 amino acids long instead of the normal 234 (Figure 3, panel B). Although Acp29AB is not detected in the mutant, *Acp29AB* mRNA levels are normal, as assessed by reverse transcriptase PCR (RT-PCR) and Northern blot analyses (supplementary material). These data suggest that the protein encoded by the *Acp29AB*^{1} allele is unstable and/or degraded. Consistent with this hypothesis, the frameshift mutation eliminates three cysteines; in other CRD-containing proteins, homologous cysteines participate in structurally important disulfide bonds (Gronwald *et al*., 1998).

The *Acp29AB*^{1} allele is likely a null mutation, since it disrupts the predicted CRD and makes no detectable Acp29AB protein. We therefore used this allele to examine the role of Acp29AB in mated females by examining post-mating phenomena in mates of *Acp29AB*^{1} males.

Acp29AB *is necessary for normal sperm storage*

In order to assess the role of Acp29AB in sperm storage, we counted sperm present in the sperm storage organs of mates of *Acp29AB*^{1} and control males. Two hours after the start of mating (ASM), we found no differences between mates of *Acp29AB*^{1} and control males in sperm stored in either the spermathecae (mean sperm stored: 241.1 vs. 235.9, respectively; two-tailed t-test: $P = 0.90$; n = 32) or the seminal receptacle (mean sperm stored: 363.6 vs. 404.2, respectively; two-tailed t-test: $P = 0.22$; n = 27), suggesting that sperm are able to enter the sperm storage organs normally. By contrast, four days ASM, we found a significant effect of male genotype on number of sperm stored in the spermathecae and in the seminal receptacle (Figure 4; Table 1). Specifically,

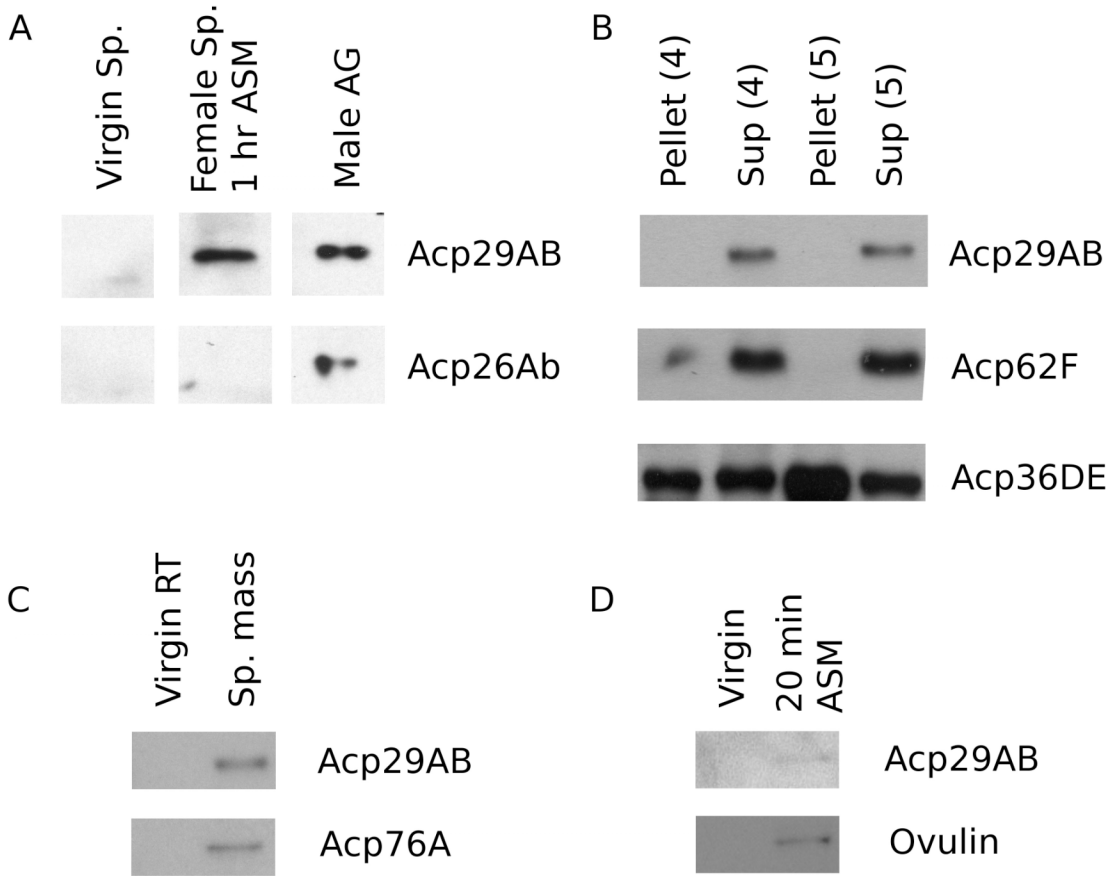


Figure 2**.** Localization of Acp29AB in the mated female. Panel A: Western blots using α-Acp29AB antibodies detect Acp29AB in the spermathecae (Sp.) of mated females 1 hour ASM, but not in the spermathecae of virgin females. Lane 3 shows accessory gland extracts from 2 males; lanes 1 and 2 each contain protein extracts from 160 spermathecae. Acp26Ab is an Acp that is known not to localize to the spermathecae (Ravi Ram and Wolfner 2005). Panel B: Acp29AB is not detectable on sperm following centrifugation. Sperm were pelleted such that sperm bound proteins remain in the pellet, with soluble proteins in the supernatant (Sup). . Acp62F is a negative control, and Acp36DE is a positive control. Panel C: Acp29AB is found in the sperm mass (Sp. mass), the mass of sperm and seminal fluid proteins transferred to females during mating. Acp76A is a positive control. Panel D: Western blots of female hemolymph. A small quantity of Acp29AB enters the hemolymph ~20 minutes ASM. Full-length Acp26Aa (shown) is known to enter the hemolymph, while the absence of low molecular weight Acp26Aa proteolysis products indicates that the hemolymph sample is not contaminated with uterine protein.

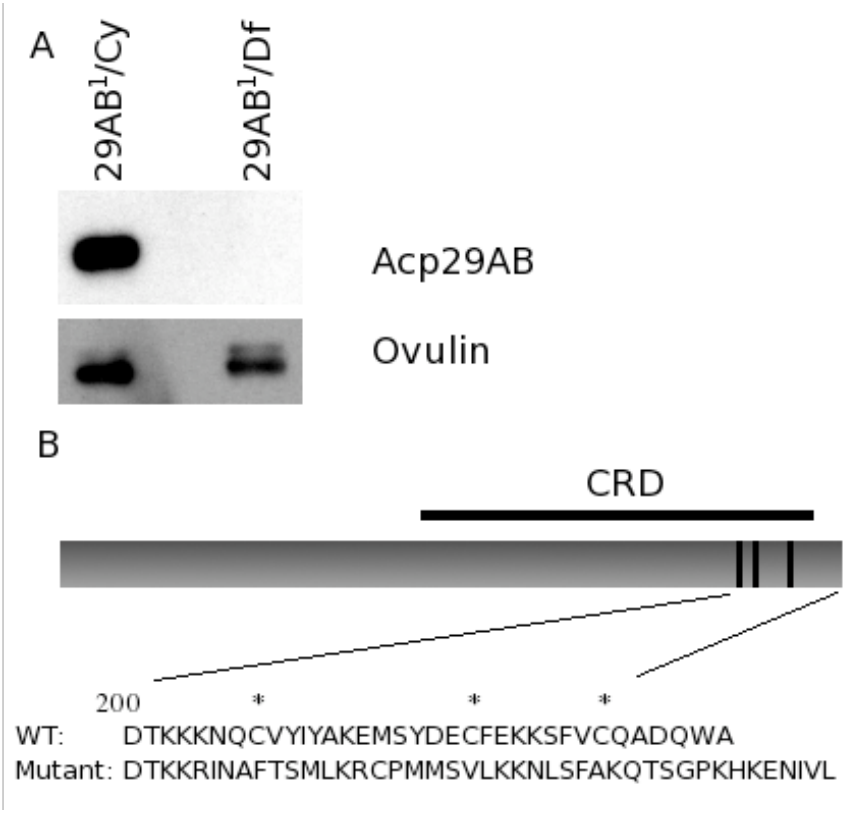


Figure 3. A frameshift mutation renders Acp29AB protein undetectable in *Acp29AB*^{1} mutants. Panel A: Anti-Acp29AB (upper panel) or anti-ovulin (lower panel) antibodies were used to probe Western blots of protein extracts from the accessory glands of Acp29AB^{1} hemizygotes or their control siblings. Panel B: Schematic representation of Acp29AB protein. The upper drawing depicts wildtype Acp29AB, showing the carbohydrate recognition domain (CRD; black bar above) and sugar binding sites (each dark line within the CRD designates 2 residues; Mueller *et al*., 2005). The predicted amino acid sequences of the C-terminus of wild-type Acp29AB, and of the *Acp29AB*^{1} mutant allele, are shown below, with putative structurally important cysteines marked by asterisks.

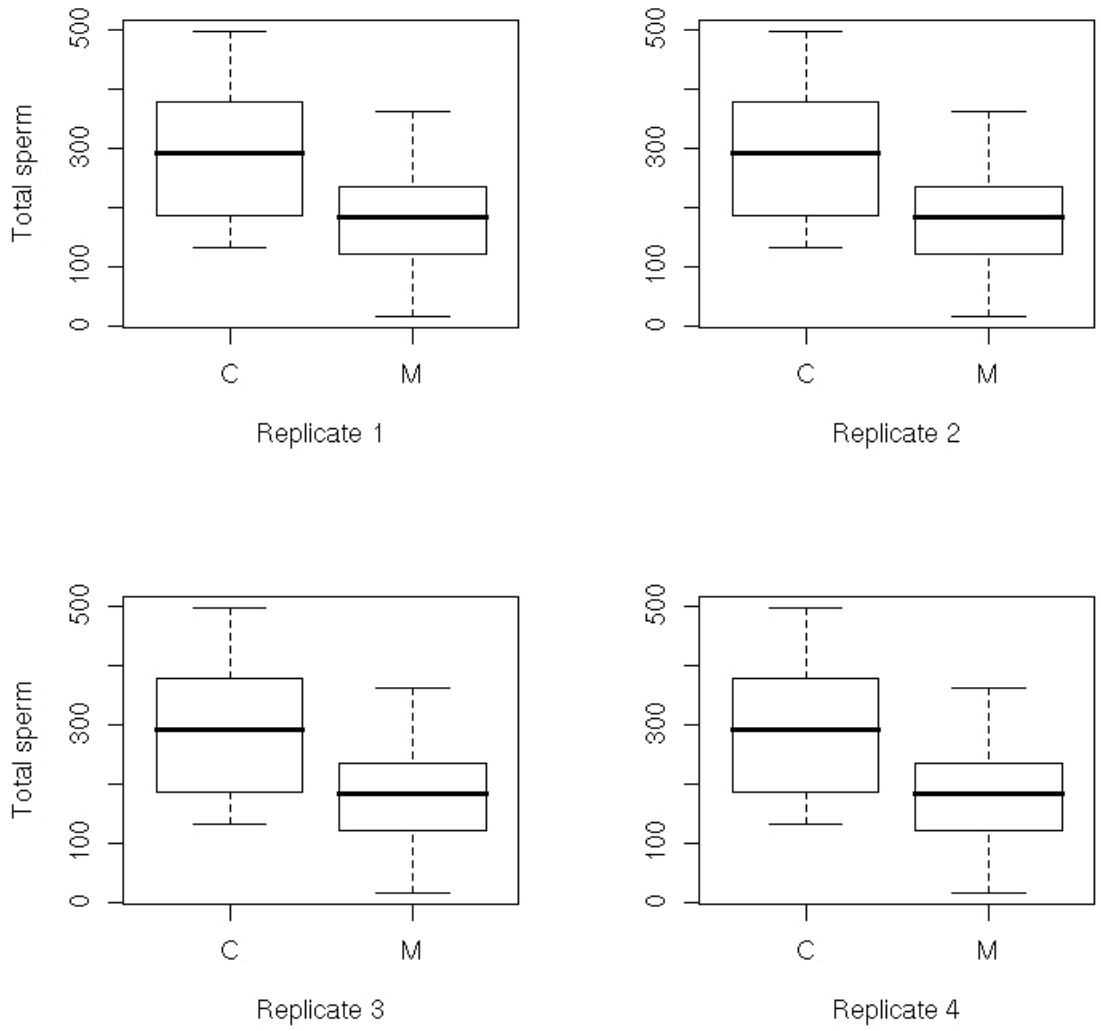


Figure 4. Reduced sperm storage 4 days after mating in mates of *Acp29AB*^{1} mutants. Total number of sperm stored by mates of control (C) or *Acp29AB*^{1} (M) males in each of four replicates is shown. In each plot, the middle horizontal line represents the median number of sperm stored, the lower and upper margins of the box represent the 25th and 75th quartiles, and the whiskers extend to 1.5 times the interquartile range from the quartiles.

Table 1: Sperm storage by *Acp29AB*[1] and control males 4 days after mating

	Effect Tests				
Tissue	Source	DF	Sum of Squares	F Ratio	*P*
Spermathecae	Genotype	1	19154.069	5.9280	0.0164
	Replicate	3	45632.342	4.7076	0.0039
	Genotype x Replicate	3	4066.584	0.4195	0.7393
Seminal Receptacle	Genotype	1	99078.50	13.7314	0.0003
	Replicate	3	300420.67	13.8785	<.0001
	Genotype x Replicate	3	48748.79	2.2520	0.0857
Combined	Genotype	1	252746.97	22.6812	<.0001
	Replicate	3	221491.23	6.6254	0.0004
	Genotype x Replicate	3	87103.30	2.6055	0.0551

fewer sperm were present in the sperm storage organs of mates of *Acp29AB*^{1} males in comparison to controls. Four independent replicate experiments were performed 4 days ASM; we found significant effects of genotype (P<0.0001) and replicate (P=0.0004) on sperm storage, but no genotype x replicate interaction effect (i.e., genotypes performed similarly in each replicate). Thus, the sperm of *Acp29AB*^{1} males do not appear to be retained efficiently in storage.

Acp29AB *may play a role in sperm competition*

We assessed the effects of the *Acp29AB*^{1} mutation on a male's sperm competitive ability, focusing on two aspects of sperm competition: P1, the proportion of offspring sired by a mutant male when he is the first of two males to mate, and P2, the proportion of offspring sired by a mutant male when he is the second of two males to mate. We found a small effect of Acp29AB on P1 (Figure 5, Panel A; Table 2): In one replicate out of three, *Acp29AB*^{1} males had a significantly reduced P1 in comparison to controls (replicate 3; P = 0.01; Mann-Whitney U-test). A similar but non-significant trend was seen in a second replicate (replicate 1; P = 0.09), and no effect was seen in a third (P = 0.66). We did not perform an analysis over all three replicates, as the combined data violate the assumption of normality typically made in ANOVA, even under several different data transformations. We failed to find an effect of Acp29AB on P2 (Table 3).

Mates of Acp29AB *deficient males do not show altered post-mating behaviors*

Stored sperm are required for several aspects of the female post-mating response, including increased egg-laying and decreased willingness to remate (Manning 1967; Chapman and Davies 2004; Ravi Ram and Wolfner 2007). Given the sperm storage phenotype of *Acp29AB*^{1} males, we tested whether mates of *Acp29AB*^{1} males showed altered egg-laying or remating behaviors. Mates of *Acp29AB*^{1} males showed no difference in remating propensity compared to controls at one or four days post-mating (Table 4). Similarly, we found no differences in total eggs laid by a female, total progeny, or egg-to-adult survivorship over ten days between mates of *Acp29AB*^{1} and control males (Table 5). Subsequent experiments focusing on late (7-10 days after mating) egg-laying similarly failed to find a significant effect of male genotype (data not shown).

Discussion

Sperm storage is vital for reproduction in many animal species, and underlies the phenomenon of sperm competition. In this study, we investigated the role of the *D. melanogaster* seminal fluid protein Acp29AB, a predicted lectin, in sperm storage. Acp29AB localizes to the female sperm storage organs following mating, with some protein also entering into the female's hemolymph. We identified a presumed loss-of-function mutation, *Acp29AB*^{1}, that disrupts the predicted carbohydrate binding domain, and that drastically reduces or eliminates the amount of Acp29AB protein present in the male accessory glands. Using this mutant, we have demonstrated that Acp29AB is necessary for the maintenance of sperm in storage at wild-type levels, although entry into storage appears to be normal. In addition, *Acp29AB*^{1} males perform poorly in the defense component of sperm competition, likely as a result of reduced numbers of stored sperm.

The Acp29AB protein is a predicted C-type lectin, and thus likely interacts with carbohydrates and/or glycoproteins in the male seminal fluid, bound to sperm, or in the female reproductive tract. Notably, protein-carbohydrate interactions play important roles in many aspects of reproduction across a wide range of animal species. Such interactions are vital for sperm-egg fusion in both invertebrates and vertebrates (reviewed in (Mengerink and Vacquier 2001); see also (Rosati et al. 2000; Intra, Cenni, and Perotti 2006), with egg glycoproteins acting as primary sperm receptors in many species. Moreover, recent studies have shown that protein-carbohydrate interactions are directly involved in the establishment and maintenance of the oviductal sperm reservoir in mammals (Suarez 2002; Ekhlasi-Hundrieser et al. 2005; Ignotz, Cho, and Suarez 2007).

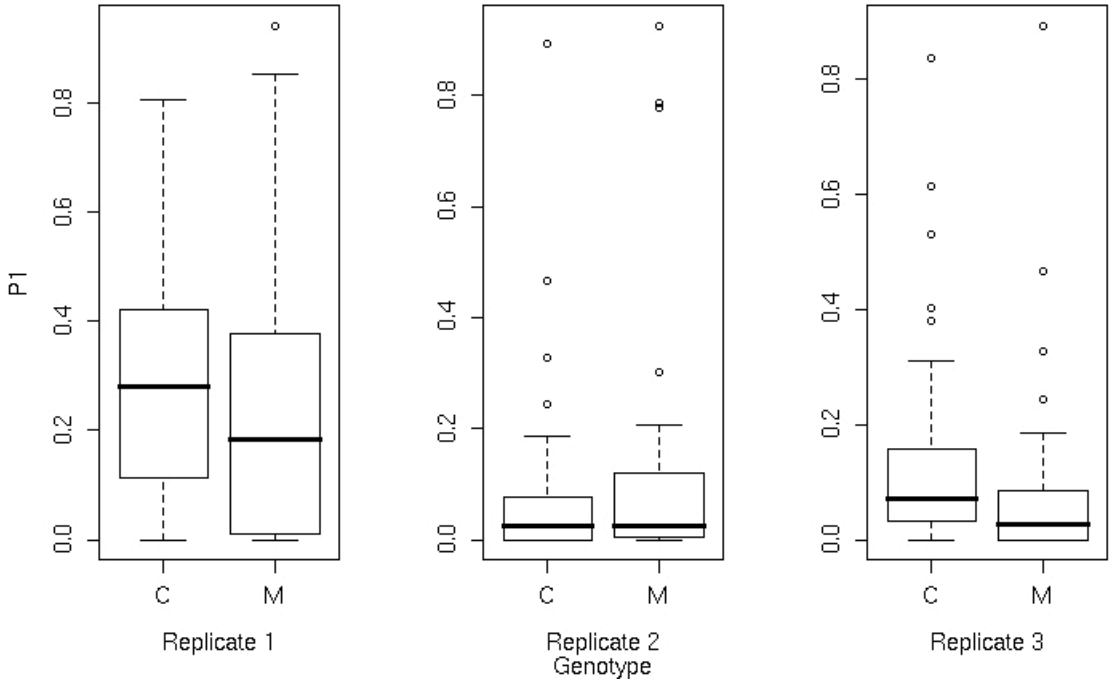


Figure 5. *Acp29AB*^{1} males perform poorly in the defense component of sperm competition. P1, the proportion of offspring sired by the first male, is given on the X-axis, for control (C) or *Acp29AB*^{1} (M) males, over three replicates. The difference between control and *Acp29AB*^{1} males is statistically significant for replicate 3 ($P = 0.01$; Mann-Whitney U-test); see Table 2. Box plots are as in Figure 4.

Table 2: Sperm defense – Proportion of offspring sired by by *Acp29AB*^{1} or control males when mating first

	Replicate 1		Replicate 2		Replicate 3	
Male genotype	n	Median P1	n	Median P1	n	Median P1
Acp29AB^{1}/Df(2L)ED611	50	0.184	44	0.024	41	0.026
Control	57	0.281	41	0.026	40	0.071
P (Mann-Whitney U)	-	0.09	-	0.66	-	.01

Table 3: Sperm offense – Proportion of offspring sired by by *Acp29AB*^{1} or control males when mating second

	Replicate 1		Replicate 2	
Male genotype	n	Median P2	n	Median P2
Acp29AB^{1}/Df(2L)ED611	43	0.903	35	0.897
Control	35	0.910	26	0.882
P (Mann-Whitney U)	-	0.91	-	0.44

Table 4: Remating behavior of mates of *Acp29AB*^{1} and control males 1 and 4 days after mating

	1 Day ASM		4 Days ASM	
First male genotype	Rep. 1	Rep. 2	Rep. 1	Rep. 2
Acp29AB^{1}/Df(2L)ED611	1/30	1/20	15/20	19/38
Control	1/26	4/20	9/18	22/36
P (Fisher's exact test)	1	0.34	0.18	0.36

Table 5: Fertility parameters of mates of *Acp29AB*^{1} and control males – totals over ten days after mating. All data were Box-Cox transformed for ANOVA to improve fit to normality.

			Effect tests (Genotype)		
Phenotype	Mutant mean (±SD)	Control mean (±SD)	Df	F-ratio	*P*
Total eggs laid	420.4 (±94.5)	452.5 (±101.9)	1	2.3877	0.13
Total progeny	346.8 (±85.8)	326.9 (±79.6)	1	1.4582	0.23
Hatchability	0.76 (±0.14)	0.79 (±0.15)	1	0.6021	0.44

.

Our finding, as well as those of Ravi Ram and Wolfner (2007), that predicted lectins are necessary for normal sperm storage and release, likewise suggests a role for protein-carbohydrate interactions in sperm storage in *Drosophila*. It is unlikely that identical mechanisms operate in mammals and flies: In cows, for example, fucose mediates the tight attachment of sperm to the oviductal epthelium (Lefebvre, Lo, and Suarez 1997), inhibiting sperm movement, whereas *D. melanogaster* sperm maintain some motility while in storage (Lefevre and Jonsson 1962). We instead propose three potential mechanisms that might explain Acp29AB's role in sperm storage: (1) Acp29AB may promote interactions between sperm and components of the lumen of the sperm storage organs that promote sperm survival and/or retention. While Acp29AB does not bind tightly to sperm (Figure 2), it does localize to the spermathecae and associates with the sperm mass. This loose association with sperm may be sufficient to mediate interactions between sperm and other molecules. (2) Acp29AB may stimulate the production or release of molecules that affect sperm storage or survival, perhaps through interaction with a glycosylated receptor. Acp29AB could exert such an effect either from within the reproductive tract, or through the neuroendocrine system, given that some Acp29AB enters the female's hemolymph. Previous experiments have suggested the existence of female-derived substances that contribute to sperm survival: As noted previously, sperm stored in the seminal receptacles of mutant females lacking spermathecae suffer from reduced viability (Anderson 1945), suggesting that the spermathecae produce viability-promoting substances. (3) Acp29AB may help to protect sperm from pathogens and/or the female's immune system.

Although the $Acp29AB^{1}$ allele's effect on sperm storage is evidently sufficient to impair a male's sperm competitive ability, we did not see an effect on other female post-mating behaviors that depend on the presence of sperm. Specifically, neither egg-laying nor re-mating propensity was affected by the presence or absence of Acp29AB in the male ejaculate. By constrast, mates of males mutant for the sperm storage protein Acp36DE do show increased re-mating propensity and decreased egg-laying. The sperm storage phenotype of mates of $Acp29AB^{1}$ males is, however, less pronounced than that of mates of *Acp36DE* mutants – lack of Acp29AB in the ejaculate leads to an approximately 40% reduction in sperm storage, whereas absence of Acp36DE leads to a ~80-90% reduction in sperm storage (Neubaum and Wolfner 1999). It is possible, then, that mates of $Acp29AB^{1}$ males have sufficient sperm in storage to manifest a normal post-mating behavioral response.

In this context, it is interesting to note that *Acp29AB* has several paralogs in the *D. melanogaster* genome, at least one of which (*lectin 29Ca*) has accessory gland specific or biased expression (Holloway and Begun 2004). The subtlety of the $Acp29AB^{1}$ mutant phenotype in comparison to that of *Acp36DE* mutants may therefore derive from functional redundancy between *Acp29AB* and at least one of its paralogs.

Several previous studies have suggested a role for Acp29AB in sperm competition, on the basis of genotype-phenotype associations. Two large association studies (Clark et al. 1995; Fiumera, Dumont, and Clark 2005) have found correlations between alleles of *Acp29AB* and sperm competitive ability – Clark et al. (1995) found an effect of *Acp29AB* genotype on P1, while Fiumera et al. (2005) found an effect on P2. We note that our failure to find an effect of *Acp29AB* genotype on P2, and Fiumera et al.'s (2005) failure to find an effect on P1, are not necessarily in conflict: The natural polymorphisms used in the association studies may not be strict loss of function alleles, and their effects likely additionally depend on female genotype (e.g. Clark, Begun, and Prout 1999). We propose that our finding, that $Acp29AB^{1}$ males suffer from reduced sperm storage, suggests a mechanism for the effects seen by Clark et al. (1995) and Fiumera et al. (2005) – differences in sperm competitive ability likely reflect differences in numbers of sperm stored, due to differences in *Acp29AB* genotype, and perhaps also due to interactions between a male's *Acp29AB* genotype and female genotype.

At least one other protein involved in sperm storage in *Drosophila* has also been implicated in sperm competition. Clark et al. (1995) found associations between sperm competitive ability and alleles of *Acp36DE*, and males null for Acp36DE perform poorly in sperm competition assays (Chapman et al. 2000). Neubaum and Wolfner (1999) and Chapman et al. (2000) argued that differences in sperm storage likely underlie differences in P1 and P2 associated with *Acp36DE* genotype. Thus, proteins involved in different stages of sperm storage (entry into storage for Acp36DE, and maintenance in storage for Acp29AB) can have similar effects during sperm competition. These data suggest that differential storage of sperm from different males due to seminal protein polymorphism may play an important role in determining the outcome of sperm competition.

Patterns of molecular evolution at *Acp29AB* are consistent with positive selection at this locus: (Aguadé 1999; Zurovcova, Tatarenkov, and Berec 2006) have found evidence for an excess of nonsynonymous (amino acid altering) substitutions between *D. melanogaster* and *D. simulans* at *Acp29AB*, suggesting multiple adaptive fixations of favorable amino acid variants between these two species. The inferred history of positive selection on *Acp29AB* may well result from its role in sperm competition, as noted by Aguadé (1999) - any variant that grants an advantage in the context of sperm competition, or that increases a sperm's probability of being used by a female, will be favored (barring antagonistic pleiotropy – see e.g., Fiumera, Dumont, and Clark 2005). Consequently, genes whose products are involved in sperm management, and hence may contribute to variation in post-copulatory sexual selection, are predicted to experience elevated rates of positive selection (e.g., Civetta and Singh 1995; Swanson and Vacquier 2002; Clark, Aagaard, and Swanson 2006).

Sperm storage is a widespread phenomenon with important functional and evolutionary consequences. Our finding, that the predicted lectin Acp29AB is necessary for the maintenance of sperm in storage, should help to elucidate mechanisms of sperm storage in *D. melanogaster*. Furthermore, our results help to establish a role for differential sperm storage in determining the outcome of sperm competition.

REFERENCES

Adams, E. M., and M. F. Wolfner. 2007. Seminal proteins but not sperm induce morphological changes in the Drosophila melanogaster female reproductive tract during sperm storage. J Insect Physiol **53**:319-331.

Aguadé, M. 1999. Positive selection drives the evolution of the Acp29AB accessory gland protein in Drosophila. Genetics **152**:543-551.

Allen, A. K., and A. C. Spradling. 2008. The Sf1-related nuclear hormone receptor Hr39 regulates Drosophila female reproductive tract development and function. Development **135**:311-321.

Bertram, M. J., D. M. Neubaum, and M. F. Wolfner. 1996. Localization of the Drosophila male accessory gland protein Acp36DE in the mated female suggests a role in sperm storage. Insect Biochem Mol Biol **26**:971-980.

Bialojan, S., D. Falkenburg, and R. Renkawitz-Pohl. 1984. Characterization and developmental expression of beta tubulin genes in Drosophila melanogaster. Embo J **3**:2543-2548.

Birkhead, T. R., and A. P. Møller. 1998. Sperm competition and sexual selection. Academic Press, San Diego, CA.

Bloch Qazi, M. C., Y. Heifetz, and M. F. Wolfner. 2003. The developments between gametogenesis and fertilization: ovulation and female sperm storage in Drosophila melanogaster. Dev Biol **256**:195-211.

Bloch Qazi, M. C., and M. F. Wolfner. 2003. An early role for the Drosophila melanogaster male seminal protein Acp36DE in female sperm storage. J Exp Biol **206**:3521-3528.

Chapman, T., and S. J. Davies. 2004. Functions and analysis of the seminal fluid proteins of male Drosophila melanogaster fruit flies. Peptides **25**:1477-1490.

Chapman, T., D. M. Neubaum, M. F. Wolfner, and L. Partridge. 2000. The role of male accessory gland protein Acp36DE in sperm competition in Drosophila melanogaster. Proc Biol Sci **267**:1097-1105.

Civetta, A., and R. S. Singh. 1995. High divergence of reproductive tract proteins and their association with postzygotic reproductive isolation in Drosophila melanogaster and Drosophila virilis group species. J Mol Evol **41**:1085-1095.

Clark, A. G., M. Aguade, T. Prout, L. G. Harshman, and C. H. Langley. 1995. Variation in sperm displacement and its association with accessory gland protein loci in Drosophila melanogaster. Genetics **139**:189-201.

Clark, A. G., D. J. Begun, and T. Prout. 1999. Female x male interactions in Drosophila sperm competition. Science **283**:217-220.

Clark, N. L., J. E. Aagaard, and W. J. Swanson. 2006. Evolution of reproductive proteins from animals and plants. Reproduction **131**:11-22.

DeMott, R. P., R. Lefebvre, and S. S. Suarez. 1995. Carbohydrates mediate the adherence of hamster sperm to oviductal epithelium. Biol Reprod **52**:1395-1403.

Eberhard, W. G. 1996. Female control: Sexual selection by cryptic female choice. Princeton University Press, Princeton, N. J.

Ekhlasi-Hundrieser, M., K. Gohr, A. Wagner, M. Tsolova, A. Petrunkina, and E. Topfer-Petersen. 2005. Spermadhesin AQN1 is a candidate receptor molecule involved in the formation of the oviductal sperm reservoir in the pig. Biol Reprod **73**:536-545.

Fiumera, A. C., B. L. Dumont, and A. G. Clark. 2005. Sperm competitive ability in Drosophila melanogaster associated with variation in male reproductive proteins. Genetics **169**:243-257.

Gilbert, D. G., and R. C. Richmond. 1981. Studies of esterase 6 in Drosophila melanogaster. VI. ejaculate competitive abilities of males having null or active alleles. Genetics **97**:85-94.

Gwathmey, T. M., G. G. Ignotz, J. L. Mueller, P. Manjunath, and S. S. Suarez. 2006. Bovine seminal plasma proteins PDC-109, BSP-A3, and BSP-30-kDa share functional roles in storing sperm in the oviduct. Biol Reprod **75**:501-507.

Holloway, A. K., and D. J. Begun. 2004. Molecular evolution and population genetics of duplicated accessory gland protein genes in Drosophila. Mol Biol Evol **21**:1625-1628.

Ignotz, G. G., M. Y. Cho, and S. S. Suarez. 2007. Annexins are candidate oviductal receptors for bovine sperm surface proteins and thus may serve to hold bovine sperm in the oviductal reservoir. Biol Reprod **77**:906-913.

Iida, K., and D. R. Cavener. 2004. Glucose dehydrogenase is required for normal sperm storage and utilization in female Drosophila melanogaster. J Exp Biol **207**:675-681.

Intra, J., F. Cenni, and M. E. Perotti. 2006. An alpha-L-fucosidase potentially involved in fertilization is present on Drosophila spermatozoa surface. Mol Reprod Dev **73**:1149-1158.

Kalb, J. M., A. J. DiBenedetto, and M. F. Wolfner. 1993. Probing the function of Drosophila melanogaster accessory glands by directed cell ablation. Proc Natl Acad Sci U S A **90**:8093-8097.

Koundakjian, E. J., D. M. Cowan, R. W. Hardy, and A. H. Becker. 2004. The Zuker collection: a resource for the analysis of autosomal gene function in Drosophila melanogaster. Genetics **167**:203-206.

Lefebvre, R., P. J. Chenoweth, M. Drost, C. T. LeClear, M. MacCubbin, J. T. Dutton, and S. S. Suarez. 1995. Characterization of the oviductal sperm reservoir in cattle. Biol Reprod **53**:1066-1074.

Lefebvre, R., M. C. Lo, and S. S. Suarez. 1997. Bovine sperm binding to oviductal epithelium involves fucose recognition. Biol Reprod **56**:1198-1204.

Lefevre, G., Jr., and U. B. Jonsson. 1962. Sperm transfer, storage, displacement, and utilization in Drosophila melanogaster. Genetics **47**:1719-1736.

Lung, O., and M. F. Wolfner. 1999. Drosophila seminal fluid proteins enter the circulatory system of the mated female fly by crossing the posterior vaginal wall. Insect Biochem Mol Biol **29**:1043-1052.

Manning, A. 1967. The control of sexual receptivity in female Drosophila. Anim Behav **15**:239-250.

Mengerink, K. J., and V. D. Vacquier. 2001. Glycobiology of sperm-egg interactions in deuterostomes. Glycobiology **11**:37R-43R.

Monsma, S. A., H. A. Harada, and M. F. Wolfner. 1990. Synthesis of two Drosophila male accessory gland proteins and their fate after transfer to the female during mating. Dev Biol **142**:465-475.

Mueller, J. L., D. R. Ripoll, C. F. Aquadro, and M. F. Wolfner. 2004. Comparative structural modeling and inference of conserved protein classes in Drosophila seminal fluid. Proc Natl Acad Sci U S A **101**:13542-13547.

Neubaum, D. M., and M. F. Wolfner. 1999. Mated Drosophila melanogaster females require a seminal fluid protein, Acp36DE, to store sperm efficiently. Genetics **153**:845-857.

Park, M., and M. F. Wolfner. 1995. Male and female cooperate in the prohormone-like processing of a Drosophila melanogaster seminal fluid protein. Dev Biol **171**:694-702.

Parker, G. A. 1998. Sperm competition and the evolution of ejaculates: towards a theory base. Pp. 3-54 *in* T. R. Birkhead, and A. P. Møller, eds. Sperm competition and sexual selection. Academic Press, San Diego.

Ram, K. R., and M. F. Wolfner. 2007. Sustained Post-Mating Response in Drosophila melanogaster Requires Multiple Seminal Fluid Proteins. PLoS Genet **3**:e238.

Ravi Ram, K., S. Ji, and M. F. Wolfner. 2005. Fates and targets of male accessory gland proteins in mated female Drosophila melanogaster. Insect Biochem Mol Biol **35**:1059-1071.

Ravi Ram, K., and M. F. Wolfner. 2007. Seminal influences: *Drosophila* Acps and the molecular interplay between males and females during reproduction. Integrative and Comparative Biology.

Rodriguez-Martinez, H. 2007. Role of the oviduct in sperm capacitation. Theriogenology **68 Suppl 1**:S138-146.

Rosati, F., A. Capone, C. D. Giovampaola, C. Brettoni, and R. Focarelli. 2000. Sperm-egg interaction at fertilization: glycans as recognition signals. Int J Dev Biol **44**:609-618.

Simmons, L. 2001. Sperm competition and its evolutionary consequences. Princeton University Press, Princeton.

Suarez, S. S. 2002. Formation of a reservoir of sperm in the oviduct. Reprod Domest Anim **37**:140-143.

Suarez, S. S., and R. A. Osman. 1987. Initiation of hyperactivated flagellar bending in mouse sperm within the female reproductive tract. Biol Reprod **36**:1191-1198.

Swanson, W. J., and V. D. Vacquier. 2002. The rapid evolution of reproductive proteins. Nat Rev Genet **3**:137-144.

Team, R. D. C. 2008. R: A language and environment for statistical computing. R Foundation for statistical computing, Vienna.

Tram, U., and M. F. Wolfner. 1998. Seminal fluid regulation of female sexual attractiveness in Drosophila melanogaster. Proc Natl Acad Sci U S A **95**:4051-4054.

Wolfner, M. F., H. A. Harada, M. J. Bertram, T. J. Stelick, K. W. Kraus, J. M. Kalb, Y. O. Lung, D. M. Neubaum, M. Park, and U. Tram. 1997. New genes for male

accessory gland proteins in Drosophila melanogaster. Insect Biochem Mol Biol **27**:825-834.

Zurovcova, M., A. Tatarenkov, and L. Berec. 2006. Differences in the pattern of evolution in six physically linked genes of Drosophila melanogaster. Gene **381**:24-33.

APPENDIX

FALSE STARTS AND LOOSE ENDS – CHARACTERIZATION OF A CANDIDATE OVULIN RECEPTOR AND A POSSIBLE REMATING MUTANT

Characterization of a candidate ovulin receptor

The identity of the receptor(s) for ovulin is currently unknown. We became interested in a G-protein coupled receptor (GPCR) located ~20 kb upstream of ovulin as a candidate receptor, given the close linkage between the ligand/receptor pair SCR and SRK in *Arabidopsis lyrata* (SCR and SRK are involved in self-incompatibility and encode a pollen ligand and stigma receptor, respectively). This GPCR, currently designated CG34381 (previously CG31645, CG14002/3/4) has no known ligand. Earlier annotations of CG34381 truncated the 5' end of the gene, removing several putative transmembrane domains. The annotation fixes this problem (verified by RT-PCR). Attempts at 5'-RACE to delimit the true 5' end of the gene have so far been unsuccessful.

I assayed eggs laid 1-day post-mating by CG34381 RNAi females or their Sb siblings. Vienna stock number 42758 was used. In two experiments, RNAi females mated to CS males laid more eggs than their Sb siblings (Figure1 1 and 2 below; Experiment 1: t-test $P = 0.030$; Experiment 2: Tukey's HSD $P = 0.0031$). As an additional control in experiment 2, I also tested egg-laying by females bearing a CG8982 (ovulin) RNAi construct, or their Sb siblings. Since ovulin is not expressed in females, no difference should have been observed. However, a difference in egg-laying comparable to that for CG34381 RNAi females was observed (Figure 2; Tukey's HSD $P = 0.000033$). Thus, CG34381's status as an ovulin receptor is still ambiguous. The experiments undertaken here should be repeated, preferably using a range of drivers and RNAi constructs (2 are at hand, VDRC numbers 7886, 42758).

Identification of a possible remating mutant

We became interested in CG13318, which encodes a putative protease homolog, because of its expression in the female reproductive tract (Swanson et al. 2004 *Genetics*) and because of evidence for positive selection in between species (PAML) analyses. I obtained a PiggyBac insertion mutant from the Bloomington stock center (stock number 10364) and a deletion covering the appropriate region (stock number 7954). 1 and 4 days after mating, insertion mutants remated significantly more often than did several controls (Tables 1 and 2; balancer siblings and a different PiggyBac insertion line – stock number 18319, insertion into the protease CG7415). However, at day 1, an extra control – a precise excision of the PiggyBac element from stock 10364 – also had a high remating frequency. This latter result suggests that the insertion into CG13318 was not responsible for the remating phenotype. Consistent with this interpretation, RNAi knockdown of CG13318 had no effect on remating in two independent lines (Table 3; lines were constructed using symPUAST by A. Wong and L. Sirot). The fact that the remating phenotype was present for the 10364/Df and excision/Df lines suggests that a mutation linked to the original PiggyBac insertion might be responsible. Of the genes deleted by the deficiency, CG11775 stood out as a promising candidate because it encodes a putative glutamate-gated ion channel. Partial sequencing of CG11775 in stock 10364 revealed a single mis-sense mutation, V87A. Phenotyping of a CG11775 RNAi line (VDRC stock number 5820) failed to replicate the remating phenotype (Table 4). Knockdown of CG11775 has not been verified in this line; this should be done before ruling out CG11775 as a candidate locus. Further deficiency mapping may help to narrow down the region responsible for the remating phenotype in stock 10364.

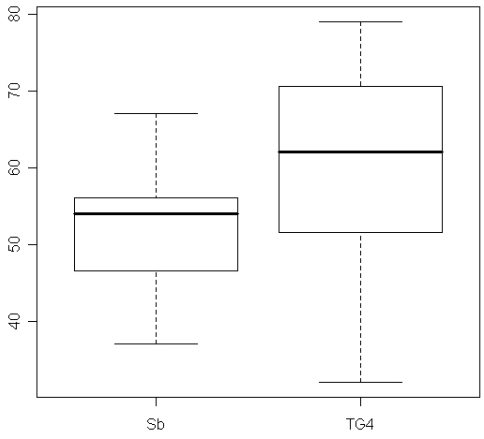


Figure 1. Eggs-laid 1 day post-mating by CG34381 RNAi ("TG4") or control ("Sb") females. 3-5 day old virgin females were mated to CS males and allowed to lay eggs on fresh food (not supplemented with additional yeast) for 24 hours. Females were aged on food supplemented with additional yeast from collection until just prior to mating.

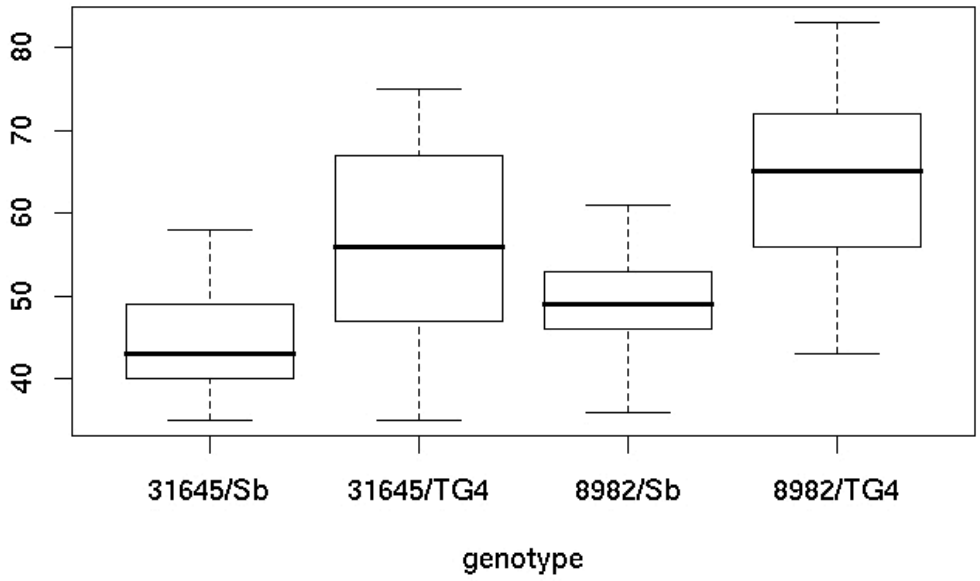


Figure 2. Eggs-laid 1 day post-mating by CG31645 (= CG34381) and CG8982 RNAi (TG4) or control (Sb) females.

Table 1: Remating at 4 days after first mating. Fisher's exact tests (two-tailed): 10364/Df vs. 10364/Hu: $P = 0.0035$; 10364/Df vs. 18319/Df: $P = 0.049$; 10364/Hu vs. 18319/Df: $P = 0.69$

Genotype	Remated	Didn't remate
10364/Df	16	3
10364/Hu (control)	4	9
18319/Df (control)	6	7

Table 2: Remating at 1 day after first mating. Fisher's exact tests (two-tailed): 10364/Df vs. 10364/Hu: $P = 0.0068$; 10364/Df vs. 18319/Df: $P = 0.017$; 10364/Df vs. Excision/Df: $P = 0.48$

Genotype	Remated	Didn't remate
10364/Df	15	4
10364/Hu (control)	5	11
18319/Df (control)	6	11
Excision/Df (control)	13	7

Table 3: Remating at 1 day after first mating by CG13318 RNAi lines

Line	Genotype	Remated	Didn't remate	*P*-value Sb vs. Tub-Gal4
1M1	Tub-Gal4	0	16	0.23
	Sb	3	15	
2M1	Tub-Gal4	1	13	1
	Sb	0	14	

Table 4: Remating 1 day after first mating by CG11775 RNAi lines. Fisher's exact test $P = 0.615$

Genotype	Remated	Didn't remate
5820/Tub-Gal4	1	24
5820/Sb	3	26

APPENDIX

LIST OF PUBLISHED WORKS

Demogines A, **Wong A**, Aquadro CF, Alani E. Accepted pending minor revisions. Incompatibilities involving yeast mismatch repair genes: a role for genetic modifiers and implications for disease penetrance and variation in genomic mutation rates. *PLoS Genet.*

Larracuente AM, Sackton TB, Greenberg AJ, **Wong A**, Singh ND, Sturgill D, Zhang Y, Oliver B, Clark AG. 2008. Evolution of protein-coding genes in Drosophila. *Trends in Genetics* 24:114-23.

Wong A, Turchin MC, Wolfner MF, Aquadro CF. 2008. Evidence for positive selection on *Drosophila melanogaster* seminal fluid protease homologs. *Molecular Biology and Evolution* 25: 497-506. Epub. 2007 Dec 4.

Drosophila 12 Genomes Consortium. 2007. Evolution of Genes and Genomes on the *Drosophila* Phylogeny. *Nature* 450: 203-18.

Haerty S, Jagadeeshan S, Kulathinal R, **Wong A**, Ravi Ram K, Sirot LK, Levesque L, Artieri C, Wolfner MF, Civetta A, Singh R. 2007. Evolution in the fast lane: rapidly evolving sex-and reproduction-related genes in species of the genus Drosophila. *Genetics* 177: 1321-35.

Williams BC, Leung G, Maiato H, **Wong A**, Li Z, Williams E, Kirkpatrick C, Aquadro C, Rieder CL, Goldberg ML. 2007. Mitch: A Rapidly Evolving Component of the Ndc80 Kinetochore Complex Required for Proper Chromosome Segregation in Drosophila. *Journal of Cell Science* 120: 3522-33.

Jensen JD, **Wong A**, Aquadro CF. 2007. On statistical and functional approaches for identifying targets of positive selection. *Trends in Genetics* 23: 568-77.

Buston PM, Bogdanowicz SM, **Wong A**, Harrison RG. 2007. Are clownfish groups composed of close relatives? An analysis of microsatellite DNA variation in *Amphiprion percula*. *Molecular Ecology* 16(17): 3671-8.

Wong A, Jensen JD, Pool JE, Aquadro CF. 2007. Phylogenetic incongruence in the *melanogaster* species group. *Mol Phy Evol* 43(3): 1138-50. Epub. 2006 Sep 9.

Wong A, Albright SN, Wolfner MF. 2006. Evidence for structural constraint on ovulin, a rapidly evolving *Drosophila melanogaster* seminal protein. *Proc Natl Acad Sci USA* 103:18644-9.

Wong A, Wolfner MF. 2006. Sexual behavior: a seminal peptide stimulates appetites. *Current Biology* 16(7):R256-7.

Pool JE, **Wong A**, Aquadro CF. 2006. Finding of male-killing *Spiroplasma* infecting *Drosophila melanogaster* in Africa implies transatlantic migration of this endosymbiont. *Heredity* 97(1): 27-32.

Swanson WJ, **Wong A**, Wolfner MF, Aquadro CF. 2004. Evolutionary EST analysis of Drosophila female reproductive tracts identifies several genes subjected to positive selection. *Genetics* 168(3): 1457-65.